

REMARKS/ARGUMENTS

Claims 58-63, 69 and 70 are pending in this application.

Claims 58-62 have been amended to remove the recitation of the phrase "native sequence." The amendments to the claims are fully supported by the specification and claims as originally filed and do not constitute new matter. Applicants believe that the current amendments place all claims in *prima facie* condition for allowance or, at least, in a better form for consideration on appeal. Accordingly, the consideration and entry of the present amendment after final rejection is respectfully requested.

Applicants expressly reserve the right to pursue any canceled matter in subsequent continuation, divisional or continuation-in-part application(s).

I. Priority

The Examiner asserts that Applicants are entitled to the priority of the filing date of the present application, October 15, 2001 allegedly "because the claimed invention is not supported by either a specific and substantial utility or a well established utility for the claimed polypeptides." (Page 3 of the instant Office Action).

As previously stated in Applicants' Responses filed on October 4, 2004, and May 23, 2005, Applicants rely on the gene amplification assay for patentable utility. The results of the gene amplification assay in lung tumors were first disclosed in U.S. Provisional Patent Application Serial No. 60/100,038, filed on September 11, 1998 and the results of the gene amplification assay in lung and colon tumors were disclosed in U.S. Provisional Patent Application Serial No. 60/131,445, filed April 28, 1999, priority to which have been claimed in this application. Accordingly, Applicants submit that the subject matter of the instant claims is supported by the disclosure in U.S. Provisional Patent Application Serial No. 60/100,038, filed on September 11, 1998 and in U.S. Provisional Patent Application Serial No. 60/131,445, filed April 28, 1999. Therefore, the effective filing date of this application is April 28, 1999, the filing date of U.S. Provisional Patent Application Serial No. 60/131,445.

II. Claim Rejections Under 35 U.S.C. §112, Second Paragraph

Claims 58-62 and 69-70 are rejected under 35 U.S.C. §112, second paragraph, as allegedly being indefinite for the recitation of a "native sequence" polypeptide. The Examiner asserts that "it is not clear how one of ordinary skill in the art would be able to determine if a sequence is 'a native sequence' or not by looking at it." (Page 3 of the instant Office Action). Without acquiescing to the PTO's arguments and solely in order to expedite prosecution of the instant application, Claims 58-62 have been amended to remove the recitation of the phrase "native sequence." Accordingly, withdrawal of the rejection under 35 U.S.C. §112, second paragraph is respectfully requested.

III. Claim Rejections Under 35 U.S.C. §§101 and 112, First Paragraph (Enablement)

Claims 58-63 and 69-70 remain rejected under 35 U.S.C. §101 allegedly "because the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility." (Page 4 of the instant Office Action).

Claims 58-63 and 69-70 further remain rejected under 35 U.S.C. §112, first paragraph, allegedly "since the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility ..., one skilled in the art clearly would not know how to use the claimed invention." (Page 4 of the instant Office Action).

For the reasons outlined below, Applicants respectfully disagree and traverse the rejection. With respect to Claims 58-63 and 69-70, Applicants submit that not only has the Patent Office not established a *prima facie* case for lack of utility and enablement, but that the PRO213-1 polypeptides possess a credible, specific and substantial asserted utility and are fully enabled.

First of all, Applicants respectfully maintain the position that the specification discloses at least one credible, substantial and specific asserted utility for the claimed PRO213-1 polypeptides for the reasons previously set forth in Applicants' Responses filed on October 4, 2004, and May 23, 2005.

Furthermore, as first discussed in Applicants' Response of October 4, 2004, Applicants rely on the gene amplification data for patentable utility of the PRO213-1 polypeptide, and the gene amplification data for the gene encoding the PRO213-1 polypeptide is clearly disclosed in

the instant specification under Example 114. As previously discussed, a ΔC_t value of at least 1.0 was observed for PRO213-1 in at least 35 of the lung and colon primary tumors and tumor cell lines listed in Table 9. Table 9 teaches that the nucleic acids encoding PRO213-1 showed 1.03 to 5.55 ΔC_t units which corresponds to $2^{1.03}$ to $2^{5.55}$ - fold amplification or 2.04 to 46.9 fold amplification in 16 different human primary lung tumors, LT1, LT1a, LT3, LT4, LT6, LT7, LT9, LT11, LT12, LT13, LT15, LT16, LT17, LT19, LT21 and LT22. PRO213-1 also showed 1.18 to 3.79 ΔC_t units which corresponds to $2^{1.18}$ to $2^{3.79}$ - fold amplification or 2.27 to 13.8 fold amplification in 11 different human primary colon tumors, CT2, CT4, CT5, CT6, CT8, CT10, CT12, CT14, CT15, CT16 and CT17. In addition, PRO213-1 showed 1.31 to 2.95 ΔC_t units which corresponds to $2^{1.31}$ to $2^{2.95}$ - fold amplification or 2.48 to 7.73 fold amplification in three different lung cancer cell lines (Calu-1, H441 and H810), and 1.22 to 2.08 ΔC_t units which corresponds to $2^{1.22}$ to $2^{2.08}$ - fold amplification or 2.33 to 4.23 fold amplification in five different colon cancer cell lines (HT29, SW403, LS174T, HCT15 and HCC2998). Accordingly, the present specification clearly discloses overwhelming evidence that the gene encoding the PRO213-1 polypeptide is significantly amplified in a significant number of lung and colon tumors.

In further support, Applicants have submitted, in their Response filed October 4, 2004, a Declaration by Dr. Audrey Goddard. Applicants particularly draw the Examiner's attention to page 3 of the Goddard Declaration which clearly states that:

It is further my considered scientific opinion that an at least **2-fold increase** in gene copy number in a tumor tissue sample relative to a normal (*i.e.*, non-tumor) sample is significant and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. Accordingly, a gene identified as being amplified at least 2-fold by the quantitative TaqMan PCR assay in a tumor sample relative to a normal sample is **useful as a marker for the diagnosis of cancer**, for monitoring cancer development and/or for measuring the efficacy of cancer therapy. (Emphasis added).

As indicated above, the gene encoding the PRO213-1 polypeptide shows at least a two fold amplification in 35 different lung and colon tumors and tumor cell lines. In addition, the Goddard Declaration clearly establishes that the TaqMan real-time PCR method described in

Example 114 has gained wide recognition for its versatility, sensitivity and accuracy, and is in extensive use for the study of gene amplification. The facts disclosed in the Declaration also confirm that based upon the gene amplification results, one of ordinary skill would find it credible that PRO213-1 is a diagnostic marker of lung and colon cancer.

The Examiner asserts that "damaged, precancerous lung epithelium is often aneuploid," and states that "[o]ne skilled in the art would not conclude that PRO213-1 is a diagnostic probe for lung cancer unless it is clear that PRO213-1 is amplified to a clearly greater extent in true lung or colon tumor tissue relative to non-cancerous lung or colon epithelium." (Page 5 of the instant Office Action). In support of this assertion the Examiner refers to the reference by Hittelman.

Applicants note that the title of the Hittelman paper is "Genetic Instabilities in Epithelial Tissues at Risk for Cancer." Hittelman studied lung tissue from chronic smokers, which had been exposed for years to carcinogenic tobacco smoke. As Hittelman explains, "[t]umors of the aerodigestive tract have been proposed to reflect a 'field cancerization' process whereby the whole tissue is exposed to carcinogenic insult (e.g., tobacco smoke) and is at increased risk for multistep tumor development (page 3). The detection of increases in chromosome number therefore identifies cells which have begun the first steps in this multistep progression to cancer. Even if these particular epithelial regions are not yet cancerous, their presence is strongly correlated with the development of cancer in the target tissue as a whole. Hittelman concludes that **"the measurement of chromosome instability in the target tissue will be useful in assessing cancer risk** as well as response to intervention" (page 10; emphasis added).

Accordingly, Hittelman shows that an increase in chromosome number or gene amplification is associated not with normal tissues, but with cancerous, or pre-cancerous tissues, and therefore, an increase in chromosome number or gene amplification is a useful marker for a cancerous or pre-cancerous state. Detection of pre-cancerous cells or tissues is useful because, as explained by Hittelman, it allows for assessing cancer risk, as well as response to intervention. Hence, Applicants respectfully submit that whether a pre-cancerous or tumor sample were analyzed, the showing of DNA amplification of the PRO213-1 gene would still be significant,

since it would lead to the diagnosis of either a pre-cancerous state or a cancerous state, which is the utility asserted here.

Despite the Examiner's assertion that such a use "is not well-established in the prior art," it is clear, as discussed above, that the use of amplified genes as markers for assessing cancer risk is explicitly contemplated in Hittelman *et al.* Further, the attached paper by Crowell *et al.* (Cancer Epidemiol. Biomarkers Prev. 5:631-637 (1996); copy enclosed as Exhibit 1) studies the detection of trisomy 7 in nonmalignant bronchial epithelium from lung patients and from individuals at high risk for lung cancer. The authors concluded that "molecular analyses may enhance the power for detecting premalignant changes in bronchial epithelium in high-risk individuals" (Abstract). Thus the use of amplified genes as markers for assessing cancer risk was explicitly contemplated in the art as early as 1996.

The Examiner asserts that the data shown in Table 9 does not provide a basis for utility or enablement of the claimed polypeptides, because "it is not predictable that gene amplification results in increased mRNA expression, or that increased mRNA expression results in increased protein production" (Page 6 of the instant Office Action). In support of this assertion, the Examiner has previously cited references by Pennica *et al.* and Gygi *et al.* The Examiner asserts that Pennica *et al.* was cited as "evidence showing a lack of correlation between gene (DNA) amplification and elevated mRNA levels." (Page 7 of the instant Office Action). Applicants respectfully submit that, for the reasons previously set forth in Applicants' Responses filed on October 4, 2004, and May 23, 2005, the teachings of Pennica *et al.* are specific to *WISP* genes, and say nothing about the correlation of gene amplification and protein expression in general. The Examiner asserts that Gygi *et al.* was cited "as providing evidence that polypeptide levels cannot be accurately predicted from mRNA levels, and that variances as much as 40-fold or 50-fold were not uncommon." (Page 7 of the instant Office Action). Yet the Examiner acknowledges that "Gygi *et al.* demonstrates that high levels of mRNA generally correlate with high levels of protein and that it appears that there is a general positive correlation between mRNA levels and protein levels." (Page 8 of the instant Office Action). Thus Gygi *et al.* supports Applicants' position that there is a positive correlation between the overexpression of mRNA and protein.

In support of the assertion that there is "a poor correlation between mRNA expression and protein abundance," the Examiner cites additional references by Lian *et al.* and Fessler *et al.* (Page 8 of the instant Office Action). The Examiner asserts that Lian *et al.* examined mRNA versus protein levels in differentiating myeloid cells and found that there was a poor correlation between mRNA expression and protein changes. Applicants submit that Lian *et al.* only teach that protein expression may not correlate with mRNA level in differentiating myeloid cells and does not teach anything regarding such a lack of correlation for genes in general. Myeloid cell differentiation relates to hematopoiesis and is an entirely different biological process from solid tumor development because these two process involve entirely different regulatory mechanisms and molecules. Analysis of surface antigens expressed on myeloid cells of the granulocyte-monocyte-histiocyte series during differentiation in normal and malignant myelomonocytic cells is useful in identifying and classifying human leukemias and lymphomas, but cannot be used in diagnosis of any solid tumors. Therefore, even if the teaching of Lian *et al.* accurately reflects the correlation between mRNA and protein for the particular system studied, it can not apply to tumor diagnosis assays of the present application.

In addition, the authors themselves admit that there are a number of problems with the data presented in this reference. At page 520 of this article, the authors explicitly express their concerns by stating that "[t]hese data must be considered with several caveats: membrane and other hydrophobic proteins and very basic proteins are not well displayed by the standard 2DE approach, and proteins presented at low level will be missed. In addition, to simplify MS analysis, we used a Coomassie dye stain rather than silver to visualize proteins, and this decreased the sensitivity of detection of minor proteins." (emphasis added). It is known in the art that Coomassie dye stain is a very insensitive method of measuring protein. This suggests that the authors relied on a very insensitive measurement of the proteins studied. The conclusions based on such measurements can hardly be accurate or generally applicable.

The Examiner also asserts that Fessler *et al.*, who examined lipopolysaccharide-activated neutrophilins, "found a 'poor concordance between mRNA transcript and protein expression changes' in human cells." (Page 6 of the instant Office Action). Again, as with Lian *et al.*, Fessler *et al.* only examined the expression level of a few proteins/RNAs in response to LPS

stimulation, which involves an entirely different regulatory mechanism from that involved in tumor development. Therefore, the teachings of Fessler *et al.* do not apply here. Additionally, the PTO has overlooked a number of limitations of the study by Fessler *et al.*

For example, as admitted by Fessler *et al.*, protein identification by two-dimensional PAGE is limited to well-resolved regions of the gel, may perform less well with hydrophobic and high molecular weight proteins, and tends to select for more abundant protein species (page 31301, col. 1). Harvesting of the LPS-incubated PMNs at 4 hours may have prevented detection of earlier, transient changes and may have thereby introduced artificial transcript-protein discordance. Furthermore, the post-LPS incubation, pre-two-dimensional PAGE cell washes would be expected to remove secreted proteins from further analysis. In addition, because protein binding of Coomassie Blue has a limited dynamic range and is typically not linear throughout the range of detection, image analysis of Coomassie Blue-stained protein spots should only be considered as semi-quantitative (see page 31301, col. 1).

In summary, both Fessler *et al.* and Lian *et al.* have relied on insensitive and inaccurate methods of measuring protein expression levels. The teachings of these two references can not be relied upon to establish a *prima facie* showing of lack of utility.

The Examiner suggests that a "very relevant reference" is Chen *et al.* (Page 8 of the instant Office Action). The Examiner cites Chen *et al.* to the effect that only twenty-eight of the 165 protein spots (17%) or 21 of 98 genes (21.4%) had a statistically significant correlation between protein and mRNA expression data. Applicants respectfully submit that the analysis by Chen *et al.* is not applicable to the present application.

First, Applicants note that proteins selected for study by Chen *et al.* were those detectable by staining of 2D gels. As noted in, for example, Haynes *et al.* (Electrophoresis 19:1862-1871 (1998); copy enclosed as Exhibit 2) there are problems with selecting proteins detectable by 2D gels. "It is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions" (page 1870, col. 1). Thus Chen *et al.* by selecting

proteins detectable by staining of 2D gels are likely to have excluded from their analysis many of the proteins most likely to be significant as cancer markers.

Secondly, Chen *et al.* looked at expression levels across a set of samples including a large number of tumor samples (76) along with a much smaller number of normal samples (9). The tumor samples were taken from stage 1 and stage III lung adenocarcinomas, which were classified as bronchoalveolar, bronchial derived or both bronchial and bronchoalveolar derived. Accordingly, the tissues examined were from different tissues in different stages of normal or cancerous growth. The authors determined the relationship between mRNA and protein expression by using the average expression values for all samples. The average value for each protein or mRNA was generated using all 85 lung tissue samples. This resulted in negative normalized protein values in some cases. Further, the authors chose an arbitrary threshold of 0.115 for the correlation to be considered significant. Accordingly, the Chen paper does not account for different expression in different tissues or different stages of cancer.

Thirdly, no attempt was made to compare expression levels in normal versus tumor samples, and in fact the authors concede that they had too few normal samples for meaningful analysis (page 310, col. 2). As a result, the analysis in the Chen paper shows only that a number of randomly selected proteins have varying degrees of correlation between mRNA and protein expression levels within a set of different lung adenocarcinoma samples. The Chen paper does not address the issue of whether increased mRNA levels in the tumor samples taken together as one group, as compared to the normal samples as a group, correlated with increased protein levels in tumorous versus normal tissue. Accordingly, the results presented in the Chen paper are not applicable to the application at issue.

The correct test of utility is whether the utility is "more likely than not". In the case of the Chen reference, even if the analysis presented is correct (which is disputed), a review of the correlation coefficient data presented in the Chen *et al.* paper indicates that it is more likely than not that increased mRNA expression correlates with increased protein expression. A review of Table 1, which lists 66 genes [the paper incorrectly states there are 69 genes listed] for which only one protein isoform is expressed, shows that 40 genes out of 66 had a positive correlation between mRNA expression and protein expression. This clearly meets the test of "more likely

than not". Similarly, in Table II, 30 genes with multiple isoforms [again the paper incorrectly states there are 29] were presented. In this case, for 22 genes out of 30, at least one isoform showed a positive correlation between mRNA expression and protein expression. Furthermore, 12 genes out of 29 showed a strong positive correlation [as determined by the authors] for at least one isoform. No genes showed a significant negative correlation. It is not surprising that not all isoforms are positively correlated with mRNA expression. Certain isoforms are likely non-functional proteins. Thus, Table II also provides that it is more likely than not that protein levels will correlate with mRNA expression levels.

The same authors in Chen *et al.*, published a later paper, Beer *et al.*, Nature Medicine 8(8) 816-824 (2002) (copy enclosed as Exhibit 3) which described gene expression of genes in adenocarcinomas and compared that to protein expression. In this paper they report that "these results suggest that the oligonucleotide microarrays provided reliable measures of gene expression" (page 817). The authors also state "these studies indicate that many of the genes identified using gene expression profiles are likely relevant to lung adenocarcinoma". Clearly the authors of the Chen paper agree that microarrays provide a reliable measure of gene expression levels and can be used to identify genes whose overexpression is associated with tumors.

Similarly, the references previously submitted by Applicants (the Orntoft, Hyman, and Pollack references), also analyzed mRNA and protein expression levels for genes known to be amplified in tumor samples. These papers also indicate that it is more likely than not that increased gene expression levels correlate with increased expression of the protein. The Chen reference does not provide sufficient evidence to dispute this finding.

The Examiner further cites Anderson *et al.* to the effect that there was a poor correlation (0.48) between mRNA and protein levels in liver cells. Applicants submit that the teachings of Anderson *et al.* do not apply to the presently claimed invention because Anderson *et al.* studied mRNA/protein correlation in proteins obtained from liver tissue, while the present invention is directed to polypeptides that are overexpressed in colon and lung tumor, which is an entirely different cellular environment from liver tissue. It would be apparent that different post-translational or post-transcriptional regulation mechanisms are involved in these two systems.

Therefore, the conclusion of Anderson *et al.* does not apply to proteins associated with tumor tissues. Moreover, even the author in this reference admitted that several experimental flaws in this paper will limit that accuracy of the data. For instant, the protein measurements rely on CBB binding and its is well-known that different proteins can bind CBB with different affinities. More significantly, the authors did not measure actual mRNA abundance for each protein, but looked at the numbers of clones found in a library. The precision of these measurements is limited because several proteins studied were represented only by one or two clones. As the authors admit, "such small numbers of clones lead to potentially large quantitative errors because of sampling error" (page 536, col. 1). As can be seen in Table 1, the data from the proteins represented only by one or two clones strongly affects the non-linearity of the total dataset. Thus these technique limitations are detrimental to the accuracy of the protein and mRNA abundance data as well as the conclusions based on these data. Finally, even assuming it is accurate, the conclusion by Anderson *et al.* does not support the Examiner's position. To the contrary, the data in Anderson *et al.* suggest that there is a significant correlation between mRNA and protein levels. Anderson *et al.* have observed a correlation coefficient of 0.48 between protein and mRNA abundance. As shown, for example, in Chen *et al.*, correlation coefficients over 0.25 are deemed to be significant (see Table II, and page 309, col. 1). In fact, the highest correlation coefficient reported by Chen *et al.* is 0.4003, less than the 0.48 observed for the Anderson *et al.* data. Accordingly, the Examiner cannot rely on the teaching of Anderson *et al.* to establish a *prima facie* showing of lack of utility.

Applicants reiterate that the evidentiary standard to be used throughout *ex parte* examination in setting forth a rejection is a preponderance of the totality of the evidence under consideration. Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Only after the Examiner has made a proper *prima facie* showing of lack of utility, does the burden of rebuttal shift to the applicant.

The Patent Office has failed to meet its initial burden of proof that Applicant's claims of utility are not substantial or credible. The arguments presented by the Examiner in combination

with the Lian *et al.*, Fessler *et al.*, Chen *et al.* and Anderson *et al.* papers do not provide sufficient reasons to doubt the statements by Applicants that PRO213-1 has utility. As set forth above, both Chen *et al.* and Anderson *et al.* support Applicants' position that there is a positive correlation between the overexpression of mRNA and protein.

In contrast, Applicants have submitted ample evidence to show that, in general, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level. First, the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.*, (made of record in Appellants' Response filed October 4, 2004) collectively teach that in general, gene amplification increases mRNA expression. Second, the Declaration of Dr. Paul Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, shows that, in general, there is a correlation between mRNA levels and polypeptide levels.

The Examiner asserts that "Orntoft *et al.* could only compare the levels of about 40 well-resolved and focused *abundant* proteins." (Page 10 of the instant Office Action; emphasis in original). While technical considerations did prevent Orntoft *et al.* from evaluating a larger number of proteins, the ones they did look at showed a clear correlation between mRNA and protein expression levels. As Orntoft *et al.* state, "In general there was a highly significant correlation ($p < 0.005$) between mRNA and protein alterations.... 26 well focused proteins whose genes had a known chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ($p < 0.005$) with the mRNA changes detected using the arrays." (See page 42, column 2 to page 34, column 2). Accordingly, Orntoft *et al.* clearly support Applicants' position that proteins expressed by genes that are amplified in tumors are useful as cancer markers.

The Examiner also appears to misunderstand the data presented by Hyman *et al.* The Examiner has asserted that "of the 12,000 transcripts analyzed, a set of 270 was identified in which overexpression was attributable to gene amplification." The Examiner concludes that "[t]his proportion is approximately 2%; the Examiner maintains that 2% does not provide a reasonable expectation that the slight amplification of PRO351 would be correlated with elevated levels of mRNA, much less protein." (Page 10 of the instant Office Action). Applicants respectfully submit that the Examiner appears to have misinterpreted the results of Hyman *et al.*

Hyman *et al.* chose to do a genome-wide analysis of a large number of genes, most of which, as shown in Figure 2, were not amplified. Accordingly, the 2% number is meaningless, as the low figure mainly results from the fact that only a small percentage of genes are amplified in the first place. The significant figure is not the percentage of genes in the genome that show amplification, but the percentage of amplified genes that demonstrate increased mRNA and protein expression.

The Examiner has further asserted that the Hyman reference "found 44% of *highly* amplified genes showing overexpression at the mRNA level, and 10.5% of *highly* overexpressed genes being amplified; thus, even at the level of high amplification and high overexpression, the two do not correlate." (Page 10 of the instant Office Action). Applicants submit that the 10.5% figure is not relevant to the issue at hand. One of skill in the art would understand that there can be more than one cause of overexpression. The issue is not whether overexpression is always, or even typically caused by gene amplification, but rather, whether gene amplification typically leads to overexpression.

The Examiner's assertion is not consistent with the interpretation Hyman *et al.* themselves place on their data, stating that, "The results illustrate **a considerable influence of copy number on gene expression patterns.**" (page 6242, col. 1; emphasis added). In the more detailed discussion of their results, Hyman *et al.* teach that "[u]p to 44% of the highly amplified transcripts (CGH ratio, >2.5) were overexpressed (*i.e.*, **belonged to the global upper 7% of expression ratios**) compared with only 6% for genes with normal copy number." (See page 6242, col. 1; emphasis added). These details make it clear that Hyman *et al.* set a highly restrictive standard for considering a gene to be overexpressed; yet almost half of all highly amplified transcripts met even this highly restrictive standard. Therefore, the analysis performed by Hyman *et al.* clearly shows that "it is more likely than not" that a gene which is amplified in tumor cells will have increased gene expression.

The Examiner further asserts that Hyman *et al.* and Pollack *et al.* do not examine protein expression. Applicants respectfully submit that the Orntoft *et al.*, Hyman *et al.* and Pollack *et al.* references were submitted primarily as evidence that in general, gene amplification increases mRNA expression. With regard to the correlation between mRNA expression and protein levels,

Applicants previously submitted a Declaration by Dr. Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, to show that mRNA expression correlates well with protein levels, in general. As previously discussed, the Utility Examination Guidelines¹ state, "Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered."

The Examiner states that in assessing the weight to be given expert testimony, "the examiner may properly consider, among other things, the nature of the fact sought to be established, the strength of any opposing evidence, the interest in the outcome of the case, and the presence or absence of factual support for the expert's opinion." (Page 11 of the instant Office Action). Applicants respectfully submit that, as discussed above, the PTO has failed to provide evidence demonstrating a lack of correlation between gene amplification and increased mRNA and protein levels, in general. Further, Dr. Polakis' statement that "an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell" is based on factual, experimental findings, clearly set forth in the Declaration. The Office Action's suggestion that Dr. Polakis might be misrepresenting these experimental results out of an interest in the outcome of the case is inappropriate.

Taken together, although there are some examples in the scientific art that do not fit within the central dogma of molecular biology that there is a correlation between polypeptide and mRNA levels, these instances are exceptions rather than the rule. In the majority of amplified genes, the teachings in the art, as exemplified by Orntoft *et al.*, Hyman *et al.*, Pollack *et al.*, and the Polakis Declaration, overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO213-1 gene, that the PRO213-1 polypeptide is concomitantly overexpressed. Thus, Applicants submit that the PRO213-1

¹ Part IIB, 66 Fed. Reg. 1098 (2001).

polypeptides have utility in the diagnosis of cancer and based on such a utility, one of skill in the art would know exactly how to use the claimed polypeptides for diagnosis of cancer.

The Examiner again cites Hu *et al.* as showing that genes displaying a 5-fold change or less in mRNA expression in tumors compared to normal showed no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease. (Page 12 of the instant Office Action).

Applicants respectfully submit that Hu *et al.* does not conclusively show that it is more likely than not that gene amplification does not result in increased expression at the mRNA and polypeptide levels. Applicants respectfully submit that Hu *et al.* manipulated various aspects of the input data in order to minimize the false positives and negatives in their analysis. Applicants further submit that the statistical analysis by Hu *et al.* is not a reliable standard because the frequency of citation only reflects the current research interest in a molecule but not the true biological function of the molecule. Finally, the conclusion in Hu *et al.* only applies to a specific type of breast tumor (estrogen receptor (ER)-positive breast tumor) and can not be generalized as a principle governing microarray study of breast cancer in general, let alone the various other types of cancer genes in general. In fact, even Hu *et al.* admit that "[i]t is likely that this threshold will change depending on the disease as well as the experiment. Interestingly, the observed correlation was only found among ER-positive (breast) tumors not ER-negative tumors." (See page 412, left column). Therefore, based on these findings, the authors add, "This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently." (*Id.*; emphasis added).

The Examiner asserts that "Applicant is holding Hu *et al.* to a higher standard than their own specification, which does not provide proper statistical analysis such as reproducibility, standard error rates, etc." (Page 13 of the instant Office Action). Applicants note that they do not argue that Hu *et al.* lacks reproducibility, standard error rates, etc. for their data, given that Hu *et al.* did a literature survey and conducted no actual experiments of their own. Rather,

Applicants' point is that, given the various biases in selecting the data to be considered, as acknowledged by the authors themselves, the collection of data surveyed by Hu *et al.* simply does not demonstrate the conclusion the PTO attempts to reach concerning a general lack of correlation between microarray data and biological significance. Accordingly, Applicants respectfully submit that the Examiner has not shown a lack of correlation between microarray data and the biological significance of cancer genes.

The Examiner has asserted that Hanna *et al.* supports the rejection, in that Hanna *et al.* "show that gene amplification does not reliably correlate with protein over-expression, and thus the level of polypeptide expression must be tested empirically." (Page 13 of the instant Office Action). Applicants respectfully point out that the Examiner appears to have misread Hanna *et al.* Hanna *et al.* clearly state that gene amplification (as measured by FISH) and polypeptide expression (as measured by immunohistochemistry, IHC) are well correlated ("in general, FISH and IHC results correlate well" (Hanna *et al.* p. 1, col. 2)). It is only a subset of tumors which show discordant results. Thus Hanna *et al.* supports Applicants' position that it is more likely than not that gene amplification correlates with increased polypeptide expression.

Applicants have clearly shown that the gene encoding the PRO213-1 polypeptide is amplified in at least 34 lung and colon tumors. Therefore, the PRO213-1 gene, similar to the HER-2/neu gene disclosed in Hanna *et al.*, is a tumor associated gene. Furthermore, as discussed above, in the majority of amplified genes, the teachings in the art overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO213-1 gene, that the PRO213-1 polypeptide is concomitantly overexpressed.

However, even if gene amplification does not result in overexpression of the gene product (*i.e.*, the protein) an analysis of the expression of the protein is useful in determining the course of treatment, as supported by the Ashkenazi Declaration and the Hanna article (submitted with Applicants' Response filed October 4, 2004). The Examiner appears to view the testing described in the Ashkenazi Declaration and the Hanna article as experiments involving further characterization of the PRO213-1 polypeptide itself. In fact, such testing is for the purpose of characterizing not the PRO213-1 polypeptide, but the tumors in which the gene encoding

PRO213-1 is amplified. The PRO213-1 polypeptides are therefore useful in tumor categorization, the results of which become an important tool in the hands of a physician enabling the selection of a treatment modality that holds the most promise for the successful treatment of a patient.

Finally, the Examiner asserts that "even if it could be established that gene amplification is reflected by increased polypeptide levels, the claims are broadly drawn to polypeptides that can be variants of the polypeptide of SEQ ID NO:506." The Examiner concludes that "such variant sequences would not reasonably be expected to show changed levels for a particular disease state." (Page 5 of the instant Office Action).

Applicants respectfully point out that the claims recite variants of SEQ ID NO:506 wherein the nucleic acid encoding said polypeptide is amplified in colon or lung tumor. Those variants whose encoding nucleic acids are not amplified in lung tumors are not encompassed by the claims. It is understood that many polypeptides and especially tumor antigens are known to have different isoforms or variants². One of skill in the art would therefore reasonably expect there to be variants of PRO213-1 that are also amplified in colon or lung tumors. The specification has provided detailed protocols for the gene amplification assay, in Example 114, such that one of ordinary skill in the art could identify those variants meeting the limitations of the claims, without any undue experimentation.

In conclusion, Applicants submit that the present rejection is based on the application of an incorrect, elevated legal standard, on misconstruction of the references and erroneous conclusions drawn there from. The issue of patentable utility should be assessed on the totality of evidence, using the preponderance evidentiary standard. It is submitted that on the totality of evidence Applicants have clearly established that the claimed invention has a substantial, specific and credible utility, for example, in the diagnosis of cancer. Further, based on this utility and the disclosure in the specification, one skilled in the art at the time the application was filed would know how to use the claimed polypeptides. Accordingly, Applicants request the Examiner

² Peng *et al.*, *Cancer Research*, 64:8911-8918 (2004); Kiss *et al.*, *Anticancer Research* 24:3965-3970 (2004); Perego *et al.*, *Molecular Carcinogenesis* 42(4):229-239 (2005); Nagao *et al.*, *Genomics* 85:462-471 (2005); Hong *et al.*, *Cancer Research* 64:5504-5510 (2004) (copies enclosed as Exhibits 4-8).

to reconsider and withdraw the rejection of Claims 58-63, 69 and 70 under 35 U.S.C. §§101 and 112.

IV. Claim Rejections Under 35 U.S.C. §112, First Paragraph (Written Description)

Claims 58-62, 69 and 70 remain rejected under 35 U.S.C. §112, first paragraph, as allegedly lacking adequate written description for the claimed variant polypeptides having at least 80-99% identity to amino acid residues 35-273 of SEQ ID NO:506, wherein the nucleic acid encoding the polypeptide is amplified in colon or lung tumors.

Applicants respectfully submit that the instant specification evidences the actual reduction to practice of the PRO213-1 polypeptide comprising amino acid residues 35-273 of SEQ ID NO:506. The Examiner has acknowledged that polypeptides comprising the sequence set forth in SEQ ID NO:506 meet the written description provision of 35 U.S.C. §112, first paragraph. (Page 13 of the Office Action mailed March 16, 2005). Thus, the genus of native sequence polypeptides with at least 80% sequence identity to amino acid residues 35-273 of SEQ ID NO:506, which possess the functional property that the nucleic acid encoding the polypeptide is amplified in colon or lung tumors, would meet the requirement of 35 U.S.C. §112, first paragraph, as providing adequate written description.

The specification describes methods for the determination of percent identity between two amino acid sequences. (See page 123, line 24 to page 125, line 14). In fact, the specification teaches specific parameters to be associated with the term "percent identity" as applied to the present invention. The specification further provides detailed guidance as to changes that may be made to a PRO polypeptide without adversely affecting its activity (page 180, line 9 to page 183, line 8). This guidance includes a listing of exemplary and preferred substitutions for each of the twenty naturally occurring amino acids (Table 6, page 182). The specification describes methods for one of ordinary skill in the art to identify polypeptides having at least 80% identity to amino acid residues 35-273 of SEQ ID NO:506 wherein the nucleic acid encoding the polypeptide is amplified in lung tumors. Example 114 of the present application provides step-by-step guidelines and protocols for the gene amplification assay. Thus one of ordinary skill in the art would have understood at the time of filing what was encompassed by the claims.

The Examiner asserts that "the skilled artisan cannot envision the *detailed chemical structure* of an encompassed polypeptide, and therefore conception is not achieved until reduction to practice has occurred, regardless of the complexity or simplicity of the method of isolation" (Page 16 of the instant Office Action; emphasis in original). In support of this assertion, the Examiner cites the cases of *Fiers v. Revel* and *Amgen v. Chugai*. (Page 16 of the instant Office Action).

Applicants submit that *Fiers v. Revel* and *Amgen v. Chugai* addressed conception and the written description requirement in the context of DNA-related inventions. The *Amgen* court held that conception of a DNA invention "has not been achieved until reduction to practice has occurred, *i.e.*, until after the gene has been isolated." 927 F.2d 1200 (Fed. Cir.), *cert. denied*, 502 U.S. 856 (1991), at 1206. The *Fiers* court extended this decision into the written description arena, holding that "[i]f a conception of a DNA requires a precise definition, such as by structure, formula, chemical name, or physical properties, as we have held, then a description also requires that degree of specificity." *Fiers*, 984 F.2d at 1171. Since the instant claims are directed to polypeptides, *Fiers* and *Amgen* are distinguished on the facts and do not apply.

More recently, in *Enzo Biochem., Inc. v. Genprobe, Inc.* 296 F.3d 1316 (Fed. Cir. 2002), the court adopted the standard that "the written description requirement can be met by 'showing that the invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics, . . . *i.e.*, complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics." *Id.* at 1324. While the invention in *Enzo* was still a DNA, the holding has been treated as being applicable to proteins as well. Indeed, the court adopted the standard from the USPTO's Written Description Examination Guidelines, which apply to both proteins and nucleic acids.

Accordingly, current applicable case law holds that biological sequences are not adequately described solely by a description of their desired functional activities. The instant claims meet the standard set by the *Enzo* court in that the claimed sequences are defined not only by functional properties, but also by structural limitations. It is well established that a combination of functional and structural features may suffice to describe a claimed genus. "An

applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics."³ As discussed above, Applicants have recited structural features, namely, 80% sequence identity to amino acid residues 35-273 of SEQ ID NO:506, which are common to the genus. The genus of claimed polypeptides is further defined by having a specific activity for the encoding nucleic acid, wherein the nucleic acid encoding the polypeptide is amplified in colon or lung tumors. Accordingly, a description of the claimed genus has been achieved.

This particular combination of functional activity and structural homology, as disclosed in the specification, has been recognized by the USPTO as sufficient to describe a claimed genus of polypeptides. The Examiner's attention is respectfully directed to Example 14 of the Synopsis of Application of Written Description Guidelines issued by the U.S. Patent Office, which clearly states that protein variants meet the requirements of 35 U.S.C. §112, first paragraph, as providing adequate written description for the claimed invention even if the specification contemplates but does not exemplify variants of the protein if (1) the procedures for making such variant proteins are routine in the art, (2) the specification provides an assay for detecting the functional activity of the protein and (3) the variant proteins possess the specified functional activity and at least 95% sequence identity to the reference sequence.

As discussed above, the procedures for making the claimed variant polypeptides are well known in the art and described in the specification. The specification also provides an assay, shown in Example 114, for detecting the recited functional activity of the nucleic acids encoding the variant polypeptides. Finally, the claimed variant polypeptides possess both the specified functional activity and a defined degree of sequence identity to the reference sequence, amino acid residues 35-273 of SEQ ID NO:506. Accordingly, the claimed polypeptide variants meet the standards set forth in the Written Description Guidelines.

³ M.P.E.P. §2163 II(A)(3)(a)

Thus the specification provides adequate written description for polypeptides having at least 80% identity to amino acid residues 35-273 of SEQ ID NO:506 wherein the nucleic acid encoding the polypeptide is amplified in lung tumors. Applicants therefore respectfully request that the Examiner reconsider and withdraw the written description rejection of Claims 58-62 and 69-70 under 35 U.S.C. §112, first paragraph.

V. Claim Rejections Under 35 U.S.C. §102 and 35 U.S.C. §103

Claims 58-63 and 69 remain rejected under 35 U.S.C. §102(e) as being anticipated by Holtzman *et al.* (U.S. Published Patent Application 20020028508), with an effective priority date of April 23, 1998. In particular, the Examiner alleges that Holtzman *et al.* disclose a protein that is 100% identical to the protein of SEQ ID NO:506. In addition, Claim 70 remains rejected under 35 U.S.C. §103(a) as being unpatentable over Holtzman *et al.* in view of Hopp *et al.*

Applicants have previously submitted a Declarations under 37 C.F.R. §1.131 by Dr. Goddard, Dr. Godowski, Dr. Gurney, Ms. Roy and Dr. Wood, that establish that Applicants had sequenced, cloned and homology human growth arrest-specific 6 (gas6) protein identified for the claimed polypeptides before April 23, 1998, which is earlier than the effective priority date of Holtzman *et al.*

The Examiner states that the Declaration filed on October 4, 2004 is unsigned. Applicants respectfully submit that copies of the Declaration signed by all of the inventors were submitted with the Preliminary Amendment filed May 23, 2005.

The Examiner states that the Declaration of Goddard, Godowski, Gurney, Roy and Wood has been considered but is ineffective to overcome the Holtzman *et al.* reference because the Holtzman *et al.* reference is a US patent application publication of an abandoned application which has a continuation that claims the same patentable invention. The Examiner states that if the reference and the instant application are commonly owned, the reference may be disqualified as prior art by an affidavit or declaration under 37 C.F.R. §1.130. The Examiner further states that if the reference and the instant application are not commonly owned, the reference can only be overcome by establishing priority of invention through interference proceedings.

Applicants submit that the reference and the instant application are not commonly owned, and thus the reference cannot be disqualified as prior art by an affidavit or declaration under 37

C.F.R. §1.130. The Declaration under 37 C.F.R. §1.131 establishes that Applicants had conceived and reduced to practice the invention before the effective filing date of the reference by Holtzman *et al.* Applicants agree that the priority of the invention can only be resolved through interference proceedings. Applicants respectfully request that this matter be held in abeyance pending the determination that the instant claims are patentable.

CONCLUSION

In conclusion, the present application is believed to be in *prima facie* condition for allowance, and an early action to that effect is respectfully solicited. Should there be any further issues outstanding, the Examiner is invited to contact the undersigned attorney at the telephone number shown below.

Although no fees are due, the Commissioner is hereby authorized to charge any fees, including any fees for extension of time, or credit overpayment to Deposit Account No. 08-1641, referencing Attorney's Docket No. 39780-2630 P1C4. Please direct any calls in connection with this application to the undersigned at the number provided below.

Respectfully submitted,

Date: November 18, 2005

By: Barrie D. Greene
Barrie D. Greene (Reg. No. 46,740)

HELLER EHRMAN LLP
275 Middlefield Road
Menlo Park, California 94025
Telephone: (650) 324-7000
Facsimile: (650) 324-0638

SV 2166516 v1
11/15/05 9:50 AM (39780.2630)

Detection of Trisomy 7 in Nonmalignant Bronchial Epithelium from Lung Cancer Patients and Individuals at Risk for Lung Cancer¹

Richard E. Crowell, Frank D. Gilliland, R. Thomas Temes, Heidi J. Harms, Robin E. Neft, Evelyn Heaphy, Dennis H. Auckley, Lida A. Crooks, Scott W. Jordan, Jonathan M. Samet, John F. Lechner, and Steven A. Belinsky²

Departments of Medicine [R. E. C., E. H., D. H. A.], Surgery [R. T. T.], and Pathology [L. A. C., S. W. J.], Albuquerque Veterans Administration Medical Center and the University of New Mexico Health Sciences Center, Albuquerque, New Mexico 87131; Inhalation Toxicology Research Institute, Albuquerque, New Mexico 87115 [H. J. H., R. E. N., J. F. L., S. A. B.]; Department of Epidemiology and Cancer Control Program, University of New Mexico Cancer Research and Treatment Center, Albuquerque, New Mexico 87131 [F. D. G.]; and Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21231 [J. M. S.]

Abstract

Early identification and subsequent intervention are needed to decrease the high mortality rate associated with lung cancer. The examination of bronchial epithelium for genetic changes could be a valuable approach to identify individuals at greatest risk. The purpose of this investigation was to assay cells recovered from nonmalignant bronchial epithelium by fluorescence *in situ* hybridization for trisomy of chromosome 7, an alteration common in non-small cell lung cancer. Bronchial epithelium was collected during bronchoscopy from 16 cigarette smokers undergoing clinical evaluation for possible lung cancer and from seven individuals with a prior history of underground uranium mining. Normal bronchial epithelium was obtained from individuals without a prior history of smoking (never smokers). Bronchial cells were collected from a segmental bronchus in up to four different lung lobes for cytology and tissue culture. Twelve of 16 smokers were diagnosed with lung cancer. Cytological changes found in bronchial epithelium included squamous metaplasia, hyperplasia, and atypical glandular cells. These changes were present in 33, 12, and 47% of sites from lung cancer patients, smokers, and former uranium miners, respectively. Less than 10% of cells recovered from the diagnostic brush had cytological changes, and in several cases, these changes were present within different lobes from the same patient. Background

frequencies for trisomy 7 were $1.4 \pm 0.3\%$ in bronchial epithelial cells from never smokers. Eighteen of 42 bronchial sites from lung cancer patients showed significantly elevated frequencies of trisomy 7 compared to never smoker controls. Six of the sites positive for trisomy 7 also contained cytological abnormalities. Trisomy 7 was found in six of seven patients diagnosed with squamous cell carcinoma, one of one patient with adenocarcinoma, but in only one of four patients with adenocarcinoma. A significant increase in trisomy 7 frequency was detected in cytologically normal bronchial epithelium collected from four sites in one cancer-free smoker, whereas epithelium from the other smokers did not contain this chromosome abnormality. Finally, trisomy 7 was observed in almost half of the former uranium miners; three of seven sites positive for trisomy 7 also exhibited hyperplasia. Two of the former uranium miners who were positive for trisomy 7 developed squamous cell carcinoma 2 years after collection of bronchial cells. To determine whether the increased frequency of trisomy 7 reflects generalized aneuploidy or specific chromosomal duplication, a subgroup of samples was evaluated for trisomy of chromosome 2; the frequency was not elevated in any of the cases as compared with controls. The studies described in this report are the first to detect and quantify the presence of trisomy 7 in subjects at risk for lung cancer. These results also demonstrate the ability to detect genetic changes in cytologically normal cells, suggesting that molecular analyses may enhance the power for detecting premalignant changes in bronchial epithelium in high-risk individuals.

Introduction

Although lung cancer is the leading cause of cancer death in the United States (1), early detection and intervention could decrease the high mortality rate associated with this disease if sensitive screening approaches could be developed (2-4). Early detection may be feasible because the entire respiratory tract is exposed to inhaled carcinogens; therefore, the whole lung is at risk for developing multiple, independently initiated sites. This "field cancerization" condition (5) is supported clinically by a high frequency of second primary tumors in lung cancer patients (6-9) and by the occurrence of progressive histological premalignant changes throughout the lower respiratory tract of cigarette smokers (10, 11). Moreover, recent studies using pathological tissues obtained after lung resection or autopsy have identified genetic aberrations associated with lung cancer in nonmalignant bronchial epithelium adjacent to tumors (12-16).

Although examination of pathological samples is useful for identifying genetic changes associated with carcinogenesis, this invasive approach for collection of clinical samples nec-

Received 1/23/96; revised 4/16/96; accepted 4/17/96.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¹ This work was supported in whole or in part by the Office of Health and Environmental Research, United States Department of Energy, under Contracts DE-AC04-76V01013 and DE-FG03-92ER61520; by NIH Grant SP50CA58184; and by the Dedicated Health Research Funds of the University of New Mexico School of Medicine.

² To whom requests for reprints should be addressed, at Inhalation Toxicology Research Institute, P. O. Box 5890, Albuquerque, NM 87185. Phone: (505) 845-1165; Fax: (505) 845-1198.

essary for early detection would not be appropriate for screening. However, bronchial epithelial cells harvested using routine clinical procedures could be examined for genetic changes as an initial approach for detecting individuals at high risk for lung cancer. This approach could also provide genetic markers for evaluating the effectiveness of chemoprevention regimens. Bronchoscopy provides direct access to viable cells within the airways and is a commonly used tool for obtaining samples from the lower respiratory tract, including bronchial epithelium (17). This procedure can be used to repeatedly sample the bronchial epithelium over time and to collect viable cells that can be expanded through tissue culture for functional assays.

Because of field cancerization, genetic abnormalities should be dispersed throughout the bronchial epithelium of persons at risk for lung cancer. The purpose of this investigation was to test this hypothesis by sampling nonmalignant bronchial epithelium from distinct locations within four different lobes of the lung from persons at risk for lung cancer and then assaying the bronchial cells for the presence of specific genetic abnormalities. Trisomy of chromosome 7 was examined in these cells, because this alteration is common in solid tumors, including lung cancer, of several different organ systems (18, 19). In addition, trisomy 7 has been detected in premalignant lesions such as villous adenoma of the colon (20), in the colonic mucosa of individuals with familial polyposis (21), and in the far margins of some resected lung tumors (22). Our results demonstrate that trisomy 7 can be detected in nonmalignant bronchial epithelium from patients with lung cancer distant to the site of the tumor and in individuals without tumors who are at high risk for lung cancer development. Together, these studies suggest that an extra copy of chromosome 7 may be an intermediate biomarker of ongoing field carcinogenesis.

Materials and Methods

Subject Recruitment. Bronchial epithelium was collected from 16 cigarette smokers undergoing a diagnostic workup for possible lung cancer and from 7 individuals with a prior history of underground uranium mining, 5 of whom were also smokers. Three individuals who had never smoked were also recruited to obtain bronchial epithelium not exposed directly to either tobacco smoke or radon progeny.

Pathology and Exposure History. Twelve of the 16 cigarette smokers who underwent diagnostic bronchoscopy were diagnosed with NSCLC.³ Seven tumors were characterized histologically as SCCs, four tumors were ACs, and one tumor was an adenosquamous cell carcinoma. Lung cancer was not evident in the other four subjects. Smoking histories ranged from 15 to 120 pack-years (defined as the number of cigarettes smoked per day times the number of years smoked). All of the former uranium miners worked underground between 2 and 20 years, with a range of 27–527 working level months. Five of the seven miners had smoking histories that ranged from 20–60 pack-years.

Bronchoscopic Collection and Processing of Bronchial Epithelium. A protocol was developed for harvesting viable bronchial epithelium from the lower respiratory tract using a standard cytology brush during bronchoscopy. After introduc-

tion into the lower respiratory tract, the bronchoscope was directed into each upper and lower lobe, and the carinal margin of a segmental orifice, usually the second and third bifurcation within the upper and lower lobes, respectively, was brushed. These sites were chosen because (a) they are high-deposition areas for particles; (b) they are associated frequently with histological changes in smokers; and (c) they represent sites where tumors commonly occur (11, 23). The area was first washed with saline to remove any nonadherent cells. Sites were not brushed if a tumor was visualized within 5 cm of the site. After brushing, the brush was withdrawn, placed in serum-free medium, and kept on ice until processed. Each site was brushed twice. The procedure was well-tolerated by all subjects, and no complications were noted related to the brushing procedure.

Bronchial cells were collected from only two of the sites in two of the subjects, from three sites in two subjects, and from all four sites in the remaining subjects. Although only two sites were brushed initially in case 1, cells were obtained from all four sites in this subject during a repeat bronchoscopy performed after the initial procedure did not yield a diagnosis. Samples were obtained from all four sites in the cancer-free current smokers and in the never smokers. In addition, bronchial epithelial cells derived at autopsy by Clonetics, Inc. (San Diego, CA) from four never smokers were also obtained to serve as additional controls. Only two sites sampled from most of the former uranium miners were available for analysis because cells recovered from the other sites had been used exclusively for cytology in another investigation.⁴

Bronchial Epithelial Cell Culture. Replicative cultures of the bronchial epithelial cells obtained by the procedure described above were established in our laboratory (24) using a serum-free medium (BEGM; Clonetics, Inc.) that is optimal for growth of these cells. Cells were removed from brushes by vigorous shaking in BEGM; cells from one brush were prepared for cytological analyses, and cells from the other brush were washed, resuspended in BEGM, seeded onto 60-mm fibronectin-coated plates, and grown at 37°C in 3% CO₂ and 21% O₂ until 80% confluence. Prior to passage, aliquots of cells were cryopreserved and stored at –145°C; other samples of cells were fixed in methanol-acetic acid (3:1). Next, the cells were washed four to six times in methanol:acetic acid and then dropped onto slides (about 2 × 10⁵ cells/slide). The effects of cell culture on the frequency of trisomy 7 in nonmalignant bronchial epithelium were examined by placing cells dispersed from brushes directly onto microscope slides followed by fixation.

Cytology. Cells from one brush from each bronchial collection site were prepared for cytological analysis by smearing the cells across a microscope slide. The cells were then fixed with 96% ethanol and stained according to the Papanicolaou procedure (25) to facilitate morphological evaluation by a cytopathologist.

Detection of Trisomy 2 and Trisomy 7. Trisomy 2 and trisomy 7 were determined by hybridization of cells with a biotinylated chromosome 2 or 7 centromere probe (Oncor; Gaithersburg, MD). The probes were denatured in hybridization buffer at 70°C for 5 min, and the slides were immersed in 70% formamide-2× SSPE at 70°C for 2 min. The probe was then applied to the slides, which were incubated in a humidified chamber at 37°C for 16 h. After incubation, the slides were washed in 0.25× SSPE (10 mM sodium phosphate monobasic monohydrate; 1 mM ethylenediamine tetraacetic acid disodium

³ The abbreviations used are: NSCLC, non-small cell lung cancer; SCC, squamous cell cancer; AC, adenocarcinoma; EGFR, epidermal growth factor receptor; FISH, fluorescence *in situ* hybridization; LOH, loss of heterozygosity; BEGM, Bronchial Epithelial Growth Medium.

⁴ Unpublished data.

salt, dihydrate; 150 mM sodium chloride, pH 7.4) for 5 min at 72°C, and the probe was detected with fluorescein-labeled avidin. Cell nuclei were visualized with propidium iodide.

Data Analysis. The number of centromeric hybridization signals in each cell were evaluated in 400 cells/slide, and the frequency of trisomy 7 on each slide was calculated by dividing the total number of cells expressing three hybridization signals by the total number of cells counted on each slide. Twenty % of the slides were scored by a second person, and frequencies for trisomy 7 differed by <0.4%. The total number of sites positive for trisomy 7 in subjects with SCC and AC were compared using Fisher's exact test.

Results

Cytology. Squamous metaplasia and atypical glandular cells, the only cytological abnormalities observed in lung cancer patients, were present in 32% of the samples (Table 1). These cytological changes were observed in <10% of the cells recovered from the diagnostic brush. Two subjects had three sites with cytological abnormalities, and five subjects had no cytological abnormalities. No samples contained tumor cells by cytology, although one of four sites in five subjects was collected from the same lobe where a tumor was later diagnosed.

Two of the 16 sites in smokers without lung cancer were cytologically abnormal (both in the same person; Table 2), whereas no atypical cells were present in the 12 sites from the three never smokers (Table 3). In former uranium miners, hyperplasia was present in bronchial cells collected from all four sites from one person, and in one site in two additional people (Table 2).

Culturing of Bronchial Epithelial Cells. The efficiency of establishing replicative cultures of the cells obtained by bronchial brushing was 100%. The serum-free medium used for these cultures is optimal for growing bronchial epithelial cells and does not support fibroblastic cell replication (25). Therefore, the cells were uniformly epitheloid in appearance. Growth potential was evaluated by passaging cells from all seven of the uranium miner cases and cases 1–6 from the lung cancer patients. Some of these cultures were maintained for up to nine passages (a minimum of 16 population doublings), and many underwent 30 divisions before senescence. However, none exhibited an indefinite population-doubling potential.

Detection of Trisomy 7 in Nonmalignant Bronchial Epithelium. Background rates of trisomy 7 were determined by examining normal human bronchial epithelial cell lines derived from autopsy cases of never smokers and bronchial epithelium collected from never smokers during bronchoscopy. In bronchial cell lines (passage 2) from four donors and bronchial epithelial cell samples obtained by bronchial brushing from the recruited never smokers (Table 3), only $1.4 \pm 0.3\%$ (SD) of the cells contained three hybridization signals for chromosome 7 with values ranging from 1 to 1.8%. These values agree with those reported by the manufacturer of the probe. Therefore, trisomy 7 frequencies of >2.0% (>2 SD above the mean for controls) were considered significantly different from controls.

Passage 1 or 2 bronchial cells from lung cancer patients were examined for trisomy 7. Eighteen of the 42 bronchial sites (43%) sampled from the 12 lung cancer patients contained trisomy 7 at frequencies ranging from 2.3 to 6.0% (Table 1; Fig. 1). Three subjects (cases 1, 2, and 11) displayed trisomy 7 in all sites collected during bronchoscopy, and in two subjects (cases 7 and 12), trisomy 7 was found in three of four sites (Table 1). Six of the 18 sites positive for trisomy 7 also contained cytologically abnormal cells. Trisomy 7 was found in six of seven

Table 1 Frequency of trisomy 7 in bronchial epithelial cells from lung cancer patients

Case	Age	Smoking (pack-yr)	Tumor diagnosis	Brush location	Cytological diagnosis	Trisomy 7 (frequency, %)
1	64	104	SCC	RLL ^a	N	2.8 ^b
				RUL	AGC	4.0 ^b
				RLL ^c	N	3.0 ^b
				RUL ^c	N	4.0 ^b
				LLL ^c	N	6.0 ^b
2	69	26	SCC	LUL ^c	SM	4.3 ^b
				RUL	SM	2.8 ^b
				LLL	SM	3.3 ^b
3	65	120	SCC	LUL	N	3.8 ^b
				RLL	AGC	2.0
				RUL	AGC	2.3 ^b
4	52	90	AC	LLL	AGC	2.0
				RLL	SM	1.5
				RUL	N	1.8
5	70	50	SCC	LLL	SM	1.5
				RLL	N	1.5
				RUL	N	1.5
6	61	93	AC	LLL	N	1.5
				RLL	N	1.5
				RUL	N	1.3
7	58	40	SCC	LLL	N	2.0
				RLL	N	1.8
				RUL	N	2.3 ^b
8	59	120	AdSCC	LLL	N	2.5 ^b
				RLL	N	2.8 ^b
				RUL	N	1.5
9	65	71	SCC	LLL	N	2.0
				RLL	AGC	2.0
				RUL	SM	2.5 ^b
10	63	45	AC	RLL	N	1.0
				RUL	N	1.8
				LLL	N	1.8
11	61	95	AC	LUL	N	1.3
				LLL	N	2.5 ^b
				RUL	N	2.8 ^b
12	76	17	SCC	LUL	N	2.0
				RLL	N	2.3 ^b
				LLL	N	2.3 ^b
				LUL	N	2.3 ^b

^aRLL, right lower lobe; RUL, right upper lobe; LLL, left lower lobe; LUL, left upper lobe; AGC, atypical glandular cells; SM, squamous metaplasia; N, normal cells; AdSCC, adenosquamous carcinoma.

^b $P < 0.05$ as compared to never-smoker controls.

^cResampled 4 months later.

patients diagnosed with SCC, whereas only one of four patients with AC displayed trisomy 7 in any site collected at bronchoscopy. Case 7, which had histological features of both SCC and AC, had one site positive for trisomy 7. The frequency of positive trisomy 7 sites in all patients with SCC within this small sample population was significantly greater than in AC patients ($P < 0.005$).

The reproducibility of detecting trisomy 7 at sites found to be positive for this abnormality was investigated in one patient (case 1) who required repeat bronchoscopy for clinical reasons. Trisomy 7 was increased similarly in the two sites brushed during both procedures, although cytological examination showed atypical cells in one site from the first bronchoscopy and cytologically normal cells from the same site collected

Table 2 Frequency of trisomy 7 in bronchial epithelial cells from cancer-free smokers and former uranium miners

Case	Age	Smoking (pack-yrs)	Radon exposure (WLMs) ^a	Brush location	Cytological diagnosis	Trisomy 7 (frequency, %)
13	81	15	0	RLL	N	1.8
				RUL	AGC	1.5
				LLL	N	1.8
14	34	24	0	LUL	SM	2.0
				RLL	N	1.3
				RUL	N	1.3
15	68	51	0	LLL	N	1.0
				LUL	N	1.3
				RLL	N	4.0 ^b
16	45	30	0	RUL	N	3.0 ^b
				LLL	N	4.3 ^b
				LUL	N	3.5 ^b
17	59	8	27	RLL	N	1.3
				RUL	N	1.5
				LLL	N	2.0
18	65	9	516	LUL	N	1.8
				LLL	N	3.0 ^b
				LUL	N	3.0 ^b
19	64	30	235	RUL	N	1.3
				LUL	N	3.3 ^b
				LLL	N	1.5
20	56	0	186	RLL	N	1.0
				LUL	N	2.0
				RLL	N	2.3 ^b
21	64	0	214	RLL	H	1.8
				LUL	N	0.8
				LLL	H	1.3
22	64	9	577	LUL	H	2.8 ^b
				RLL	H	2.5 ^b
				LLL	H	3.3 ^b
23	67	31	124	LUL	H	1.3
				RLL	H	2.8 ^b
				LLL	H	2.5 ^b

^a Abbreviations are as indicated in Table 1 footnote. WLM, working level month; H, hyperplasia.

^b $P < 0.05$ as compared to never-smoker controls.

during the second procedure (Table 1). The other two sites collected during the second bronchoscopy also showed elevated frequencies of trisomy 7 in this patient.

Trisomy 7 was detected in cytologically normal bronchial epithelium collected from four sites in one (case 15) of the cancer-free smokers (Table 2). Bronchial cells from the other smokers did not contain this chromosome abnormality. In the former uranium miners (cases 17–23), seven of 15 sites collected during bronchoscopy were positive for trisomy 7. Three of the positive sites were found in one subject (case 23) and also contained basal cell hyperplasia. However, the other four samples positive for trisomy 7 showed no cytological abnormality.

Two of the former uranium miners (cases 18 and 23) developed lung cancer within 2 years of bronchial cell collection. SCC was diagnosed in the right upper lobe of both subjects. As noted in Table 2, both cases were positive for trisomy 7 in the right upper lobe brushing site obtained at the initial bronchoscopy.

Tissue Culture Effects on Trisomy 7 Expression in Bronchial Epithelium. The effect of tissue culture on trisomy 7 frequency was assessed by comparing the frequency of this chromosome abnormality in freshly isolated bronchial epithelium obtained directly from bronchial brushes ("preculture") to passage 1 cells. This comparison was conducted on cells collected from two different bronchial sites in three different subjects [(cases 11 and 16 and donor 7 (never smoker)]. Cultured samples positive for trisomy 7 in case 11 were also

Table 3 Interphase analysis of chromosome 7 in normal human bronchial epithelial cells

Bronchial epithelial cell lines were established from never smokers (Clonetics) after autopsy and from volunteers. The normal distribution of chromosome 7 copy number as detected by FISH is shown by the percentage of cells exhibiting 1, 2, 3, or 4 hybridization signals. Four hundred cells containing hybridization signal were counted per donor.

Donor	Age	Brush location	Number of hybridization signals/cell (%)			
			1	2	3	4
1	6	NA ^a	3.5	92.0	1.5	3.0
2	17	NA	2.3	95.5	1.3	1.0
3	15	NA	1.5	94.7	1.8	2.0
4	41	NA	2.0	94.8	1.0	2.3
5	45	RLL	1.0	95.5	1.8	1.7
		RUL	0.5	98.3	1.0	0.2
		LLL	1.3	96.5	1.0	1.2
6	35	LUL	1.0	96.3	1.2	1.5
		RLL	1.0	96.8	1.0	1.2
		RUL	2.5	93.3	1.7	2.5
7	33	LLL	2.0	94.8	1.5	1.7
		LUL	1.8	94.2	1.8	2.2
		RLL	0.5	98.2	0.8	0.5
		RUL	0.5	97.2	1.3	1.0
		LLL	1.2	96.8	1.3	0.7
		LUL	1.0	96.0	1.5	1.5

^a Abbreviations are as indicated in the legend to Table 1. NA, not applicable.

positive in preculture cells from the same bronchial collection site, whereas sites negative for trisomy 7 in cultured cells from case 16 and the never smoker were also negative in preculture cells (data not shown). Values for trisomy 7 differed by $<0.3\%$ between preculture and cultured cells. The effect of passaging cells on the frequency of trisomy 7 was also examined in bronchial cells from case 1. Trisomy 7 frequency was similar in cells from passages 1, 4, and 7.

Frequency of Trisomy 2 in Nonmalignant Bronchial Epithelium. Aneuploidy has been detected in bronchial squamous metaplasia, a likely precursor to SCC (26). To determine whether the increased frequency of trisomy 7 detected in the current study reflects generalized aneuploidy or a specific chromosomal duplication, a subgroup of samples was evaluated for trisomy of chromosome 2. The frequency of trisomy 2 in never smokers was $1.5 \pm 0.4\%$ (data not shown). Bronchial cells from eight subjects, six of whom had elevated frequencies for trisomy 7, were evaluated. The frequency for trisomy of chromosome 2 did not differ from never smokers (Table 4).

Discussion

The studies described in this report are the first to detect and quantify an increase in trisomy 7 in the airway cells of subjects at risk for lung cancer. The presence of trisomy 7 appeared to be a specific chromosome gain and not due to generalized aneuploidy in these cells. In addition, trisomy 7 in nonmalignant epithelium from lung cancer patients was associated with SCC tumor histology, suggesting that patients with this genetic change may be at greater risk for developing SCC than other histological forms of lung cancer. This supposition was supported by the fact that two cancer-free former uranium miners with bronchial cells positive for trisomy 7 ultimately developed SCC. Finally, these results demonstrate the ability to detect genetic changes in cytologically normal cells, suggesting that molecular analyses may enhance the power for detecting

Fig. 1. FISH for chromosome 7 in bronchial epithelial cells. Trisomy 7 is apparent in one cell from this field. Magnification, $\times 530$.



Table 4 Frequency of trisomy 2 in bronchial epithelial cells from lung cancer patients, cancer-free smokers, and former uranium miners

Case	Tumor diagnosis	Brush location	Trisomy 2 (frequency, %)
1	SCC	RLL*	1.5
		RUL	1.8
		LLL	1.8
		LUL	1.0
2	SCC	LLL	1.0
		LUL	1.0
7	SCC	RLL	1.5
		RUL	2.1
		LLL	1.8
		LUL	1.5
8	AC	RLL	0.3
		RUL	1.5
		LLL	0.8
13	None	RLL	1.0
		RUL	0.8
		LLL	1.0
		LUL	1.3
15	None	RLL	1.8
		RUL	2.0
		LLL	1.0
		LUL	1.3
19	None	LUL	1.9
		RLL	0.8
23	None	RLL	1.5

* Abbreviations are as indicated in legend to Table 1.

premalignant changes in bronchial epithelium in high-risk individuals.

Cigarette smoking and the exposure of underground miners to radon progeny are both well-established respiratory carcinogens (27, 28). Tobacco smoke contains numerous mutagens and carcinogens, and radon progeny that have been inhaled and deposited on the respiratory epithelium release α

particles capable of damaging DNA (28). Although comparison between findings in the cigarette smokers and the former uranium miners is constrained by the number of participants in the two groups, trisomy 7 was found in both groups. These results are consistent with the synergism between smoking and radon progeny, which suggests commonality in the pathways by which the two carcinogens cause lung cancer (29).

The bronchial brushing method used for collecting cells from the lower respiratory tract is rapid (10–12 min total for two brushes at four different sites), well tolerated by the patient, and permits collection of viable bronchial cells that can be expanded through tissue culture at 100% efficiency. The stability of these cells in culture was evident by the fact that the frequency of trisomy 7 did not differ between primary brush cells and cells propagated for up to seven passages. Furthermore, this procedure is amenable to the production of sufficient cell numbers (1×10^6) at low passage (one or two) to accommodate multiple molecular analyses. Although the media used in culturing of bronchial epithelial cells did not appear to provide a selective growth advantage to cells harboring an additional chromosome 7, the modulation of medium supplements might lead to the establishment of clonal populations of premalignant cells. Such cell populations would greatly facilitate the identification of additional early gene changes in respiratory carcinogenesis.

The detection of trisomy 7 in multiple nonmalignant sites within the bronchial tree supports the theory of field cancerization (5), which states that diffuse exposure of the entire respiratory tract to inhaled carcinogens causes the development of multiple, independently initiated sites that can lead to tumor development. Although the frequency of this chromosome abnormality was relatively low (2.3–6.0%), these values were consistent with the low percentage of cells within each brush sample (10%) that exhibited abnormal cytology. These results are also similar to studies of chromosome gain in patients with head and neck cancer where trisomy 7 was detected at frequen-

cies of 2, 3, and 21% in histologically normal, hyperplastic, and dysplastic cells, respectively (30).

The detection of trisomy 7 in normal, hyperplastic, and metaplastic bronchial epithelium from cancer-free patients extends a recent report describing LOH at chromosomes 3p, 5q, and 9p in dysplastic premalignant bronchial lesions harvested from current and former smokers by bronchoscopy (31). The inability to detect LOH at these chromosome loci in normal or early premalignant epithelium may stem from a difference in sensitivity between the methodologies used. The low frequency of trisomy 7 and cytologically abnormal cells collected from bronchoscopy is consistent with a lack of clonality within the brush cells. FISH assays on interphase cells permit screening of individual cells, and sensitivity for detection is limited only by the number of cells examined. In contrast, microsatellite analyses for LOH cannot detect nonclonal changes but require that the chromosome alteration be present in approximately 40–50% of the sample (32, 33).

The role of trisomy 7 in lung cancer development has not been elucidated. Increased expression of EGFR, which is located on chromosome 7 (34), is observed in 50–80% of NSCLCs (16, 35, 36). EGFR expression appears greater in SCC than AC (35, 36) and is amplified in some cell lines derived from SCC (37). These findings corroborate our hypothesis that acquisition of trisomy 7 in bronchial epithelium could be prognostic for development of SCC. Moreover, expression of this gene is also increased in nonmalignant bronchial epithelium from NSCLC patients (16, 35) and in normal or premalignant epithelium adjacent to head and neck tumors (38). Thus, altered expression of EGFR could enable cells that have acquired additional genetic changes to proliferate continually and escape from terminal differentiation (39). In addition, the *c-met* oncogene is also located on chromosome 7 and is overexpressed in NSCLCs (40, 41). This oncogene encodes a transmembrane tyrosine kinase (42) that functions as a receptor for the hepatocyte growth factor (43) and is involved in sustaining the growth of NSCLC cells in culture (44).

Previous studies have detected mutations in p53 (12, 14, 35), chromosome losses at 9p21 (45) and 3p (46) in preinvasive bronchial lesions, and simple chromosome rearrangements in normal bronchial epithelium from proximal airways (47) of lung cancer patients. The prevalence of these genetic changes in normal epithelium from persons at risk for lung cancer should be quantified by FISH to define the temporal sequences of somatic genetic changes that precede the development of clonal lesions in the lung. This information will be invaluable in providing biological markers that can qualitatively estimate the extent of field cancerization in persons at risk for lung cancer and can be used to assess the efficacy of chemoprevention trials. Ultimately, the efficiency for detecting these biological markers in bronchial epithelium *versus* exfoliated epithelial cells within sputum must be established to support the use of a "genetic-based" screening approach for individuals at high risk for lung cancer. The results of the current investigation have identified one potential biomarker, trisomy 7, that may be useful in early detection and intervention for lung carcinogenesis.

References

- Boring, C. C., Squires, T. S., and Tong, T. Cancer statistics, 1993. *J. Clin. Oncol.* 11: 7–26, 1993.
- Lippman, S. M., Benner, S. E., and Hong, W. K. Cancer chemoprevention. *J. Clin. Oncol.* 12: 851–873, 1994.
- Lippman, S. M., and Spitz, M. R. Intervention in the premalignant process. *Cancer Bull.* 43: 473–474, 1991.
- Berlin, N. I., Buncher, C. R., Fontana, R. S., Frost, J. K., and Melamed, M. R. The National Cancer Institute cooperative early lung cancer detection program. Early lung cancer detection. *Am. Rev. Respir. Dis.* 130: 545–570, 1984.
- Slaughter, D. P., Southwick, H. W., and Smejkal, W. Field cancerization in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer (Phila.)* 5: 963–968, 1953.
- van Bodegom, P. C., Wagenaar, S. S., Corrin, B., Baak, J. P., Berkel, J., and Vanderschueren, R. G. Second primary lung cancer: importance of long term follow-up. *Thorax* 44: 788–793, 1989.
- Boice, J. D., and Fraumeni, J. F. Second cancer following cancer of the respiratory system in Connecticut, 1935–1982. *J. Natl. Cancer Inst. Monogr.* 68: 83–98, 1985.
- Pairolero, P. C., Williams, D. E., Bergstralh, E. J., Pichler, J. M., Bernatz, P. E., and Payne, N. J. Postsurgical stage I bronchogenic carcinoma. Morbid implications of recurrent disease. *Ann. Thorac. Surg.* 38: 331–338, 1984.
- Shields, T. W., Humphrey, E. W., Higgins, G. A., and Keehn, R. J. Long term survivors after resection of lung carcinoma. *J. Thorac. Cardiovasc. Surg.* 76: 439–442, 1978.
- Auerbach, O., Stout, A. P., Hammond, E. C., and Garfinkel, L. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *N. Engl. J. Med.* 276: 111–118, 1962.
- Auerbach, O., Hammond, E. C., and Garfinkel, L. Changes in bronchial epithelium in relation to cigarette smoking, 1955–1960 vs. 1970–1977. *N. Engl. J. Med.* 300: 381–386, 1979.
- Sundaresan, V., Ganly, P., Hasleton, P., Rudd, R., Sinha, G., Bleehen, N. M., and Rabbitts, P. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumors, are detectable in preinvasive lesions of the bronchus. *Oncogene* 7: 1989–1997, 1992.
- Sozzi, G., Miozzo, M., Donghi, R., Pilotti, S., Cariani, C. T., Pastorino, U., Pianta, G. P., and Pierotti, M. A. Deletions of 17p and p53 mutations in preneoplastic lesions of the lung. *Cancer Res.* 52: 6079–6082, 1992.
- Bennett, W. P., Colby, T. V., Travis, W. D., Borkowski, A., Jones, R. T., Lane, D. P., Metcalf, R. A., Samet, J. M., Takeshima, Y., Gu, J. R., Vähäkangas, K. H., Soini, N., Pääkkö, P., Welsh, J. A., Trump, B. F., and Harris, C. C. p53 protein accumulates frequently in early bronchial neoplasia. *Cancer Res.* 53: 4817–4822, 1993.
- Sozzi, G., Miozzo, M., Pastorino, U., Pilotti, S., Donghi, R., Giarola, M., Gregorio, L. D., Manenti, G., Radice, P., Minoletti, F., Porta, G. D., and Pierotti, M. A. Genetic evidence for an independent origin of multiple preneoplastic and neoplastic lung lesions. *Cancer Res.* 55: 135–149, 1995.
- Sozzi, G., Miozzo, M., Tagliabue, E., Calderone, C., Lombardi, L., Pilotti, S., Pastorino, U., Pierotti, M. A., and Porta, G. D. Cytogenetic abnormalities and overexpression of receptors for growth factors in normal bronchial epithelium and tumor samples of lung cancer patients. *Cancer Res.* 51: 400–404, 1991.
- Campbell, A. M., Chavez, P., Vignola, A. M., Bousquet, J., Couret, I., Michel, F. B., and Godard, P. H. Functional characteristics of bronchial epithelium obtained by brushing from asthmatic and normal subjects. *Am. Rev. Respir. Dis.* 147: 529–534, 1993.
- Testa, J., and Siegfried, J. M. Chromosome abnormalities in human non-small cell lung cancer. *Cancer Res.* 52 (Suppl.): 2702–2706, 1992.
- Matturri, L., and Lavezzi, A. M. Recurrent chromosome alterations in non-small cell lung cancer. *Eur. J. Histochem.* 38: 53–58, 1994.
- Reichmann, A., Martin, P., and Levin, B. Karyotypic findings in a colonic villous adenoma. *Cancer Genet. Cytogenet.* 7: 51–57, 1982.
- Moertel, C. A., DeWald, G. W., Coffey, R. J., and Gordon, H. Cytogenetic examination of colonic mucosa in familial polyposis. In: *Proceedings of the Second International Conference on Chromosomes in Solid Tumors*, Tucson, Arizona Cancer Center, pp. 41–48. University of Arizona, 1987.
- Lee, J. S., Pathak, S., Hopwood, V., Tomasovic, B., Mullins, T. D., Baker, F. L., Spitzer, G., and Neidhart, J. A. Involvement of chromosome 7 in primary lung tumor and nonmalignant normal lung tissue. *Cancer Res.* 47: 6349–6352, 1987.
- Ishikawa, Y., Nakagawa, K., Satoh, Y., Kitagawa, T., Sugano, H., Hirano, T., and Tsuchiya, E. Hot spots of chromium accumulation at bifurcations of chromate workers' bronchi. *Cancer Res.* 54: 2342–2346, 1994.
- Lechner, J. F., and LaVeck, M. A. A serum-free method for culturing normal human bronchial epithelial cells at clonal density. *J. Tissue Culture Methods* 9: 43–48, 1985.
- Saccomanno, G. *Pulmonary Cytology*, Ed. 2. Chicago: American Society of Clinical Pathologists Press, 1986.
- Lee, J. S., Lippman, S. M., Hong, W. K., Ro, J. Y., Kim, S. Y., Lotan, R., and Hittelman, W. N. Determination of biomarkers for intermediate end points in chemoprevention. *Cancer Res.* 52 (Suppl.): 2702S–2710S, 1992.
- United States Department of Health and Human Services. Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General. Department of Health and Human Services Publication No. (CDC)

- 89-8411. Washington, DC: United States Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 1989.
28. National Research Council. Report of the Committee on the Biological Effects of Ionizing Radiation: Health Effects of Radon and Other Internally Deposited α Emitters (BEIR IV). Washington DC: National Academy Press, 1988.
29. Lubin, J. H., Boice, J. D., Jr., Edling, C., Hornung, R. W., Howe, G. R., Kunz, E., Kusiak, R. A., Morrison, H. I., Radford, E. P., Samet, J. M., Tirmarche, M., Woodward, A., Yao, S. X., and Pierce, D. A. Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *J. Natl. Cancer Inst.*, 87: 817-827, 1995.
30. Voravud, N., Shin, D. M., Ro, J. Y., Lee, J. S., Hong, W. K., and Hittelman, W. N. Increased polysomies of chromosomes 7 and 17 during head and neck multistage tumorigenesis. *Cancer Res.*, 53: 2874-2883, 1993.
31. Thiberville, L., Payne, P., Vielkinds, J., LeRiche, J., Horsman, D., Nouvet, G., and Palcic, B. Evidence of cumulative gene losses with progression of premalignant epithelial lesions to carcinoma of the bronchus. *Cancer Res.*, 55: 5133-5139, 1995.
32. Shiseki, M., Kohno, T., Nishikawa, R., Sameshima, Y., Mizoguchi, H., and Yokota, J. Frequent allelic loss on chromosomes 2q, 18q, and 22q in advanced non-small cell lung carcinoma. *Cancer Res.*, 54: 5643-5648, 1994.
33. Merlo, A., Mabry, M., Gabrielson, E., Vollmer, R., Baylin, S. B., and Sidransky, D. Frequent microsatellite instability in primary small cell lung cancer. *Cancer Res.*, 54: 2098-2101, 1994.
34. Spurr, N. K., Solomon, E., Jansson, M., Sheen, D., Goodfellow, P. N., Bodmer, W. F., and Verström, B. Chromosomal localization of the human homologues to the oncogenes *erb-A* and *B*. *EMBO J.*, 3: 159-164, 1984.
35. Rusch, V., Klimstra, D., Linkov, I., and Dmitrovsky, E. Aberrant expression of p53 or the epidermal growth factor receptor is frequent in early bronchial neoplasia, and coexpression precedes squamous cell carcinoma development. *Cancer Res.*, 55: 1365-1372, 1995.
36. Veale, D., Ashcroft, T., March, C., Gibson, G. J., and Harris, A. L. Epidermal growth factor receptors in non-small cell lung cancer. *Br. J. Cancer*, 55: 513-516, 1987.
37. Tadashi, Y., Kamata, N., Kawano, H., Shimizu, S., Kuroki, T., Toyoshima, K., Rikimura, K., Nomura, N., Ishizaki, R., Pastan, I., Gambou, J., and Shimizu, N. High incidence of amplification of the epidermal growth factor receptor gene in human squamous carcinoma cell lines. *Cancer Res.*, 46: 414-416, 1986.
38. Shin, D. M., Ro, J. Y., Hong, W. K., and Hittelman, W. N. Dysregulation of epidermal growth factor receptor expression in premalignant lesions during head and neck tumorigenesis. *Cancer Res.*, 54: 3153-3159, 1994.
39. Soschek, C. M., and King, L. E. Functional and structural characteristics of EGF and its receptor and their relationship to transforming proteins. *J. Cell. Biochem.*, 31: 135-152, 1986.
40. Prat, M., Narsimhan, R. P., Crepaldi, T., Nicotra, M. R., Natali, P. G., and Comoglio, P. M. The receptor encoded by the human *c-met* oncogene is expressed in hepatocytes, epithelial cells, and solid tumors. *Int. J. Cancer*, 49: 323-328, 1991.
41. Liu, C., and Tsao, M-S. *In vitro* and *in vivo* expression of transforming growth factor α and tyrosine kinase receptors in human non-small cell lung carcinoma cell lines. *Am. J. Pathol.*, 142: 1155-1162, 1993.
42. Giordano, S., Ponzetto, C., Di Renzo, M. F., Cooper, S., and Comoglio, P. M. Tyrosine kinase receptor indistinguishable from the *c-met* protein. *Nature (Lond.)*, 339: 155-156, 1989.
43. Naldini, L., Vigna, E., Narsimhan, R. P., Gaudino, G., Zarnegar, R., Michalopoulos, G. K., and Comoglio, P. M. Hepatocyte growth factor (HGF) stimulates the tyrosine kinase activity of the receptor encoded by the proto-oncogene *c-MET*. *Oncogene*, 6: 501-504, 1991.
44. Liu, C., and Tsao, M-S. Proto-oncogene and growth factor/receptor expression in the establishment of primary human non-small cell lung carcinoma cell lines. *Am. J. Pathol.*, 142: 413-423, 1991.
45. Kishimoto, Y., Sugio, K., Hung, J. Y., Virmani, A. K., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific loss in chromosome 9p loci in preneoplastic lesions accompanying non-small-cell lung cancers. *J. Natl. Cancer Inst.*, 87: 1224-1229, 1995.
46. Hung, J., Kishimoto, Y., Sugio, K., Virmani, A., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific chromosome 3p deletions occur at an early stage in the pathogenesis of lung carcinoma. *JAMA*, 273: 558-563, 1995.
47. Pastorino, U., Sozzi, G., Miozzo, M., Tagliabue, E., Pilotti, S., and Pierotti, M. A. Genetic changes in lung cancer. *J. Cell. Biochem.*, 17F (Suppl.): 237-248, 1993.

Review

Paul A. Haynes
Steven P. Gyi
Daniel Flgeys
Ruedi Aebersold

Department of Molecular
Biotechnology, University of
Washington, Seattle, WA, USA

Proteome analysis: Biological assay or data archive?

In this review we examine the current state of proteome analysis. There are three main issues discussed: why it is necessary to study proteomes; how proteomes can be analyzed with current technology; and how proteome analysis can be used to enhance biological research. We conclude that proteome analysis is an essential tool in the understanding of regulated biological systems. Current technology, while still mostly limited to the more abundant proteins, enables the use of proteome analysis both to establish databases of proteins present, and to perform biological assays involving measurement of multiple variables. We believe that the utility of proteome analysis in future biological research will continue to be enhanced by further improvements in analytical technology.

Contents

1	Introduction	1862
2	Rationale for proteome analysis	1862
2.1	Correlation between mRNA and protein expression levels	1863
2.2	Proteins are dynamically modified and processed	1863
2.3	Proteomes are dynamic and reflect the state of a biological system	1863
3	Description and assessment of current proteome analysis technology	1863
3.1	Technical requirements of proteome technology	1863
3.2	2D electrophoresis - mass spectrometry: a common implementation of proteome analysis	1864
3.3	Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS	1865
3.3.1	LC-MS/MS	1865
3.3.2	Capillary LC-MS	1865
3.3.3	CE-MS/MS	1865
3.4	Assessment of 2-DE-MS proteome technology	1866
4	Utility of proteome analysis for biological research	1868
4.1	The proteome as a database	1868
4.2	The proteome as a biological assay	1868
5	Concluding remarks	1870
6	References	1870

1 Introduction

A proteome has been defined as the protein complement expressed by the genome of an organism, or, in multicellular organisms, as the protein complement expressed by a tissue or differentiated cell [1]. In the most common implementation of proteome analysis the proteins extracted from the cell or tissue analyzed are separated by high

resolution two-dimensional gel electrophoresis (2-DE), detected in the gel and identified by their amino acid sequence. The ease, sensitivity and speed with which gel-separated proteins can be identified by the use of recently developed mass spectrometric techniques have dramatically increased the interest in proteome technology. One of the most attractive features of such analyses is that complex biological systems can potentially be studied in their entirety, rather than as a multitude of individual components. This makes it far easier to uncover the many complex, and often obscure, relationships between mature gene products in cells. Large-scale proteome characterization projects have been undertaken for a number of different organisms and cell types: Microbial proteome projects currently in progress include, for example: *Saccharomyces cerevisiae* [2], *Salmonella enterica* [3], *Spiroplasma melliferum* [4], *Mycobacterium tuberculosis* [5], *Ochrobactrum anthropi* [6], *Haemophilus influenzae* [7], *Synechocystis* spp. [8], *Escherichia coli* [9], *Rhizobium leguminosarum* [10], and *Dictyostelium discoideum* [11]. Proteome projects underway for tissues of more complex organisms include those for: human bladder squamous cell carcinomas [12], human liver [13], human plasma [13], human keratinocytes [12], human fibroblasts [12], mouse kidney [12], and rat serum [14]. In this manuscript we critically assess the concept of proteome analysis and the technical feasibility of establishing complete proteome maps, and discuss ways in which proteome analysis and biological research intersect.

2 Rationale for proteome analysis

The dramatic growth in both the number of genome projects and the speed with which genome sequences are being determined has generated huge amounts of sequence information, for some species even complete genomic sequences ([15-17]). The description of the state of a biological system by the quantitative measurement of system components has long been a primary objective in molecular biology. With recent technical advances including the development of differential display-PCR [18], cDNA microarray and DNA chip technology [19, 20] and serial analysis of gene expression (SAGE) [21, 22], it is now feasible to establish global and quantitative mRNA expression maps of cells and tissues, in which the sequence of all the genes is known, at a speed and sensitivity which is not matched by current

Correspondence: Professor Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA, 98195, USA (Tel: +206-685-4235; Fax: +206-685-6392; E-mail: ruedi@u.washington.edu)

Abbreviations: CID, collision-induced dissociation; MS/MS, tandem mass spectrometry; SAGE, serial analysis of gene expression

Keywords: Proteome / Two-dimensional polyacrylamide gel electrophoresis / Tandem mass spectrometry

protein analysis technology. Given the long-standing paradigm in biology that DNA synthesizes RNA which synthesizes protein, and the ability to rapidly establish global, quantitative mRNA expression maps, the questions which arise are why technically complex proteome projects should be undertaken and what specific types of information could be expected from proteome projects which cannot be obtained from genomic and transcript profiling projects. We see three main reasons for proteome analysis to become an essential component in the comprehensive analysis of biological systems. (i) Protein expression levels are not predictable from the mRNA expression levels, (ii) proteins are dynamically modified and processed in ways which are not necessarily apparent from the gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

2.1 Correlation between mRNA and protein expression levels

Interpretations of quantitative mRNA expression profiles frequently implicitly or explicitly assume that for specific genes the transcript levels are indicative of the levels of protein expression. As part of an ongoing study in our laboratory, we have determined the correlation of expression at the mRNA and protein levels for a population of selected genes in the yeast *Saccharomyces cerevisiae* growing at mid-log phase (S. P. Gygi *et al.*, submitted for publication). mRNA expression levels were calculated from published SAGE frequency tables [22]. Protein expression levels were quantified by metabolic radiolabeling of the yeast proteins, liquid scintillation counting of the protein spots separated by high resolution 2-DE and mass spectrometric identification of the protein(s) migrating to each spot. The selected 80 samples constitute a relatively homogeneous group with respect to predicted half-life and expression level of the protein products. Thus far, we have found a general trend but no strong correlation between protein and transcript levels (Fig. 1). For some genes studied equivalent mRNA transcript levels translated into protein abundances which varied by more than 50-fold. Similarly, equivalent steady-state protein expression levels were maintained by transcript levels varying by as much as 40-fold (S. P. Gygi *et al.*, submitted). These results suggest that even for a population of genes predicted to be relatively homogeneous with respect to protein half-life and gene expression, the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript.

2.2 Proteins are dynamically modified and processed

In the mature, biologically active form many proteins are post-translationally modified by glycosylation, phosphorylation, prenylation, acylation, ubiquitination or one or more of many other modifications [23] and many proteins are only functional if specifically associated or complexed with other molecules, including DNA, RNA, proteins and organic and inorganic cofactors. Frequently, modifications are dynamic and reversible and may alter the precise three-dimensional structure and the state of activity of a protein. Collectively, the state of modification of the proteins which constitute a biological system

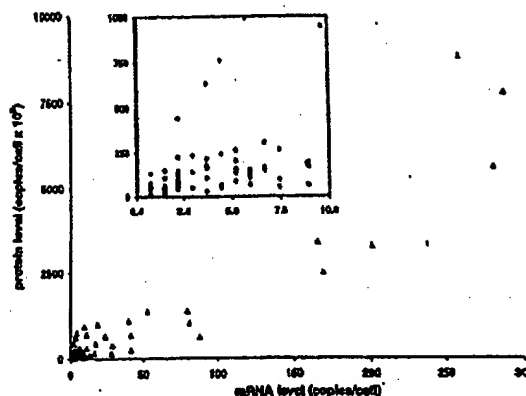


Figure 1. Correlation between mRNA and protein levels in yeast cells. For a selected population of 80 genes, protein levels were measured by 35 S-radiolabeling and mRNA levels were calculated from published SAGE tables. Inset: expanded view of the low abundance region. For more experimental details, also see Figs. 5 and 6, (S. P. Gygi *et al.*, submitted).

are important indicators for the state of the system. The type of protein modification and the sites modified at a specific cellular state can usually not be determined from the gene sequence alone.

2.3 Proteomes are dynamic and reflect the state of a biological system

A single genome can give rise to many qualitatively and quantitatively different proteomes. Specific stages of the cell cycle and states of differentiation, responses to growth and nutrient conditions, temperature and stress, and pathological conditions represent cellular states which are characterized by significantly different proteomes. The proteome, in principle, also reflects events that are under translational and post-translational control. It is therefore expected that proteomics will be able to provide the most precise and detailed molecular description of the state of a cell or tissue, provided that the external conditions defining the state are carefully determined. In answer to the question of whether the study of proteomes is necessary for the analysis of biomolecular systems, it is evident that the analysis of mature protein products in cells is essential as there are numerous levels of control of protein synthesis, degradation, processing and modification, which are only apparent by direct protein analysis.

3 Description and assessment of current proteome analysis technology

3.1 Technical requirements of proteome technology

In biological systems the level of expression as well as the states of modification, processing and macro-molecular association of proteins are controlled and modulated depending on the state of the system. Comprehensive analysis of the identity, quantity and state of modification of proteins therefore requires the detection and

quantitation of the proteins which constitute the system, and analysis of differentially processed forms. There are a number of inherent difficulties in protein analysis which complicate these tasks. First, proteins cannot be amplified. It is possible to produce large amounts of a particular protein by over-expression in specific cell systems. However, since many proteins are dynamically post-translationally modified, they cannot be easily amplified in the form in which they finally function in the biological system. It is frequently difficult to purify from the native source sufficient amounts of a protein for analysis. From a technological point of view this translates into the need for high sensitivity analytical techniques. Second, many proteins are modified and processed post-translationally. Therefore, in addition to the protein identity, the structural basis for differentially modified isoforms also needs to be determined. The distribution of a constant amount of protein over several differentially modified isoforms further reduces the amount of each species available for analysis. The complexity and dynamics of post-translational protein editing thus significantly complicates proteome studies. Third, proteins vary dramatically with respect to their solubility in commonly used solvents. There are few, if any, solvent conditions in which all proteins are soluble and which are also compatible with protein analysis. This makes the development of protein purification methods particularly difficult since both protein purification and solubility have to be achieved under the same conditions. Detergents, in particular sodium dodecyl sulfate (SDS), are frequently added to aqueous solvents to maintain protein solubility. The compatibility with SDS is a big advantage of SDS polyacrylamide gel electrophoresis (SDS-PAGE) over other protein separation techniques. Thus, SDS-PAGE and two-dimensional gel electrophoresis, which also uses SDS and other detergents, are the most general and preferred methods for the purification of small amounts of proteins, provided that activity does not necessarily need to be maintained. Lastly, the number of proteins in a given cell system is typically in the thousands. Any attempt to identify and categorize all of these must use methods which are as rapid as possible to allow completion of the project within a reasonable time frame. Therefore, a successful, general proteomics technology requires high sensitivity, high throughput, the ability to differentiate differentially modified proteins, and the ability to quantitatively display and analyze all the proteins present in a sample.

3.2 2-D electrophoresis - mass spectrometry: a common implementation of proteome analysis

The most common currently used implementation of proteome analysis technology is based on the separation of proteins by two-dimensional (IEF/SDS-PAGE) gel electrophoresis and their subsequent identification and analysis by mass spectrometry (MS) or tandem mass spectrometry (MS/MS). In 2-DE, proteins are first separated by isoelectric focusing (IEF) and then by SDS-PAGE, in the second, perpendicular dimension. Separated proteins are visualized at high sensitivity by staining or autoradiography, producing two-dimensional arrays of proteins. 2-DE gels are, at present, the most commonly used means of global display of proteins in complex

samples. The separation of thousands of proteins has been achieved in a single gel [24, 25] and differentially modified proteins are frequently separated. Due to the compatibility of 2-DE with high concentrations of detergents, protein denaturants and other additives promoting protein solubility, the technique is widely used.

The second step of this type of proteome analysis is the identification and analysis of separated proteins. Individual proteins from polyacrylamide gels have traditionally been identified using *N*-terminal sequencing [26, 27], internal peptide sequencing [28, 29], immunoblotting or comigration with known proteins [30]. The recent dramatic growth of large-scale genomic and expressed sequence tag (EST) sequence databases has resulted in a fundamental change in the way proteins are identified by their amino acid sequence. Rather than by the traditional methods described above, protein sequences are now frequently determined by correlating mass spectral or tandem mass spectral data of peptides derived from proteins, with the information contained in sequence databases [31-33].

There are a number of alternative approaches to proteome analysis currently under development. There is considerable interest in developing a proteome analysis strategy which bypasses 2-DE altogether, because it is considered a relatively slow and tedious process, and because of perceived difficulties in extracting proteins from the gel matrix for analysis. However, 2-DE as a starting point for proteome analysis has many advantages compared to other techniques available today. The most significant strengths of the 2-DE-MS approach include the relatively uniform behavior of proteins in gels, the ability to quantify spots and the high resolution and simultaneous display of hundreds to thousands of proteins within a reasonable time frame.

A schematic diagram of a typical procedure of the identification of gel-separated proteins is shown in Fig. 2. Protein spots detected in the gel are enzymatically or chemically fragmented and the peptide fragments are isolated for analysis, as already indicated, most frequently by MS or MS/MS. There are numerous protocols for the generation of peptide fragments from gel-separated proteins. They can be grouped into two categories, digestion in the gel slice [28, 34] or digestion after electrotransfer out of the gel onto a suitable membrane [29, 35-37] and reviewed in [38]). In most instances either technique is applicable and yields good results. The analysis of MS or MS/MS data is an important step in the whole process because MS instruments can generate an enormous amount of information which cannot easily be managed manually. Recently, a number of groups have developed software systems dedicated to the use of peptide MS and MS/MS spectra for the identification of proteins. Proteins are identified by correlating the information contained in the MS spectra of protein digests or MS/MS spectra of individual peptides with data contained in DNA or protein sequence databases.

The systems we are currently using in our laboratory are based on the separation of the peptides contained in protein digests by narrow bore or capillary liquid chromatog-

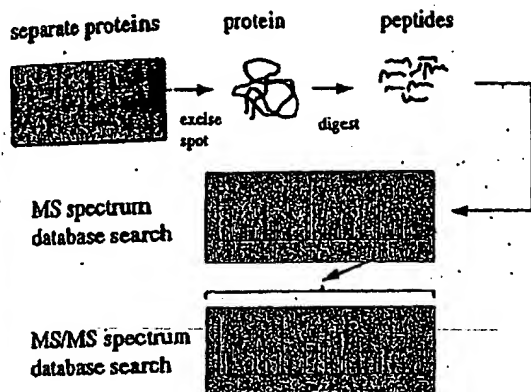


Figure 2. Schematic diagram of a procedure for identification of gel-separated proteins. Peptides can either be separated by a technique such as LC or CE, or infused as a mixture and sorted in the MS. Database searching can either be performed on peptide masses from an MS spectrum, peptide fragment masses from CID spectra of peptides, or a combination of both.

raphy [39, 40] or capillary electrophoresis [41], the analysis of the separated peptides by electrospray ionization (ESI) MS/MS, and the correlation of the generated peptide spectra with sequence databases using the SEQUEST program developed at the University of Washington [32, 33]. The system automatically performs the following operations: a particular peptide ion characterized by its mass-to-charge ratio is selected in the MS out of all the peptide ions present in the system at a particular time; the selected peptide ion is collided in a collision cell with argon (collision-induced dissociation, CID) and the masses of the resulting fragment ions are determined in the second sector of the tandem MS; this experimentally determined CID spectrum is then correlated with the CID spectra predicted from all the peptides in a sequence database which have essentially the same mass as the peptide selected for CID; this correlation matches the isolated peptide with a sequence segment in a database and thus identifies the protein from which the peptide was derived. There are a number of alternative programs which use peptide CID spectra for protein identification, but we use the SEQUEST system because it is currently the most highly automated program and has proven to be successful, versatile and robust.

3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS

It has been demonstrated repeatedly that MS has a very high intrinsic sensitivity. For the routine analysis of gel-separated proteins at high sensitivity, the most significant challenge is the handling of small amounts of sample. The crux of the problem is the extraction and transfer of peptide mixtures generated by the digestion of low nanogram amounts of protein, from gels into the MS/MS system without significant loss of sample or introduction of unwanted contaminants. We employ three different systems for introducing gel-purified samples into an MS, depending on the level of sensitivity

required. As an approximate guideline, for samples containing tens of picomoles of peptides, LC-MS/MS is most appropriate; for samples containing low picomole amounts to high femtomole amounts we use capillary LC-MS/MS; and for samples containing femtomoles or less, CE-MS/MS is the method of choice.

3.3.1 LC-MS/MS

The coupling of an MS to an HPLC system using a 0.5 mm diameter or bigger reverse phase (RP) column has been described in detail [42]. This system has several advantages if a large number of samples are to be analyzed and all are available in sufficient quantity. The LC-MS and database searching program can be run in a fully automated mode using an autosampler, thus maximizing sample throughput and minimizing the need for operator interference. The relatively large column is tolerant of high levels of impurities from either gel preparation or sample matrix. Lastly, if configured with a flow-splitter and micro-sprayer [40], analyses can be performed on a small fraction of the sample (less than 5%) while the remainder of the sample is recovered in very pure solvents. This latter feature is particularly useful when an orthogonal technique is also used to analyze peptide fractions, such as scintillation of an introduced radiolabel, and this data can be correlated with peptides identified by CID spectra.

3.3.2 Capillary LC-MS

An increase of sensitivity of approximately tenfold can be achieved by using a capillary LC system with a 100 μ m ID column rather than a 0.5 mm ID column as referred to above. Since very low flow rates are required for such columns, most reports have used a precolumn flow splitting system for producing solvent gradients. We have recently described the design and construction of a novel gradient mixing system which enables the formation of reproducible gradients at very low flow rates (low nL/min) without the need for flow splitting (A. Ducret *et al.*, submitted for publication). Using this capillary LC-MS/MS system we were able to identify gel-separated proteins if low picomole to high femtomole amounts were loaded onto the gel [40]. This system is as yet not automated and, like all capillary LC systems, is prone to blockage of the columns by microparticulates when analyzing gel-separated proteins.

3.3.3 CE-MS/MS

The highest level of sensitivity for analyzing gel-separated proteins can be achieved by using capillary electrophoresis – mass spectrometry (CE-MS). We have described in the past a solid-phase extraction capillary electrophoresis (SPE-CE) system which was used with triple quadrupole and ion trap ESI-MS/MS systems for the identification of proteins at the low femtomole to sub-femtomole sensitivity level [43, 44]. While this system is highly sensitive, its operation is labor-intensive and its operation has not been automated. In order to devise an analytical system with both the sensitivity of a CE and the level of automation of LC, we have constructed

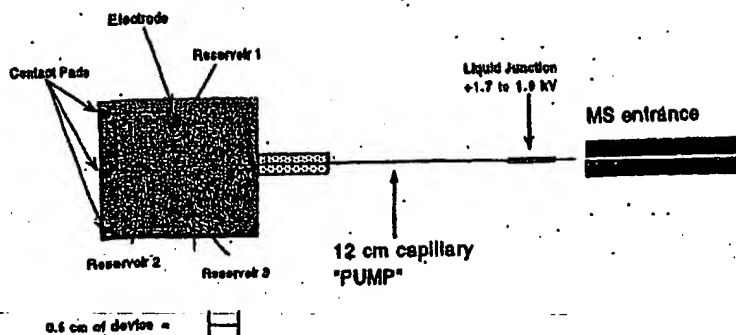


Figure 3. Schematic illustration of a microfabricated analytical system for CE, consisting of a micromachined device, coated capillary electroosmotic pump, and microelectrospray interface. The dimensions of the channels and reservoir are as indicated in the text. The channels on the device were graphically enhanced to make them more visible. Reproduced from [45], with permission.

microfabricated devices for the introduction of samples into ESI-MS for high-sensitivity peptide analysis.

The basic device is a piece of glass into which channels of 10–30 μm in depth and 50–70 μm in diameter are etched by using photolithography/etching techniques similar to the ones used in the semiconductor industry. (A simple device is shown in Fig. 3). The channels are connected to an external high voltage power supply [45]. Samples are manipulated on the device and off the device to the MS by applying different potentials to the reservoirs. This creates a solvent flow by electroosmotic pumping which can be redirected by changing the position of the electrode. Therefore, without the need for valves or gates and without any external pumping, the flow can be redirected by simply switching the position of the electrodes on the device. The direction and rate of the flow can be modulated by the size and the polarity of the electric field applied and also by the charge state of the surface.

The type of data generated by the system is illustrated in Fig. 4, which shows the mass spectrum of a peptide sample representing the tryptic digest of carbonic anhydrase at 290 fmol/ μL . Each numbered peak indicates a peptide successfully identified as being derived from carbonic an-

hydrase. Some of the unassigned signals may be chemical or peptide contaminants. The MS is programmed to automatically select each peak and subject the peptide to CID. The resulting CID spectra are then used to identify the protein by correlation with sequence databases. Therefore, this system allows us to concurrently apply a number of protein digests onto the device, to sequentially mobilize the samples, to automatically generate CID spectra of selected peptide ions and to search sequence databases for protein identification. These steps are performed automatically without the need for user input and proteins can be identified at very low femtomole level sensitivity at a rate of approximately one protein per 15 min.

3.4 Assessment of 2-DE-MS proteome technology

Using a combination of the analytical techniques described above we have identified the 80 protein spots indicated in Fig. 5. The protein pattern was generated by separating a total of 40 microgram of protein contained in a total cell lysate of the yeast strain YPH499 by high resolution 2-DE and silver staining of the separated proteins. To estimate how far this type of proteome analysis can penetrate towards the identification of low abundance proteins, we have calculated the codon bias of the genes encoding the respective proteins. Codon bias is a

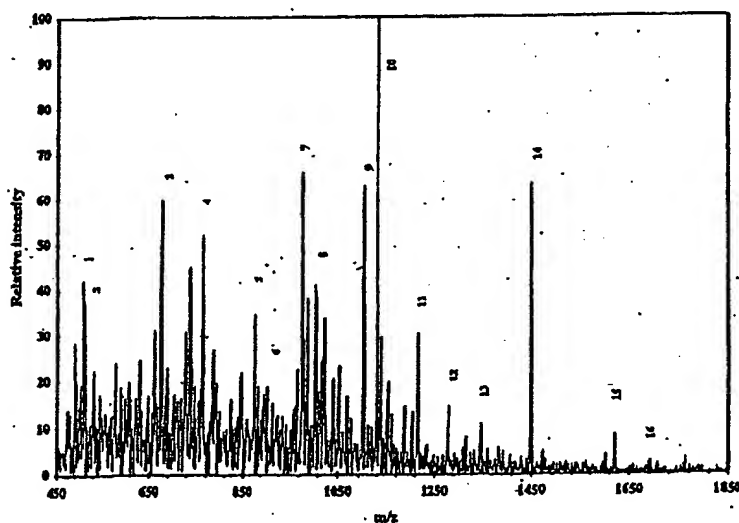


Figure 4. MS spectrum of a tryptic digest of carbonic anhydrase using the microfabricated system shown in Fig. 3. 290 fmol/ μL of carbonic anhydrase tryptic digest was infused into a Finalgan LCQ ion trap MS. Each peak was selected for CID, and those which were identified as containing peptides derived from carbonic anhydrase are numbered. Reproduced from [45], with permission.

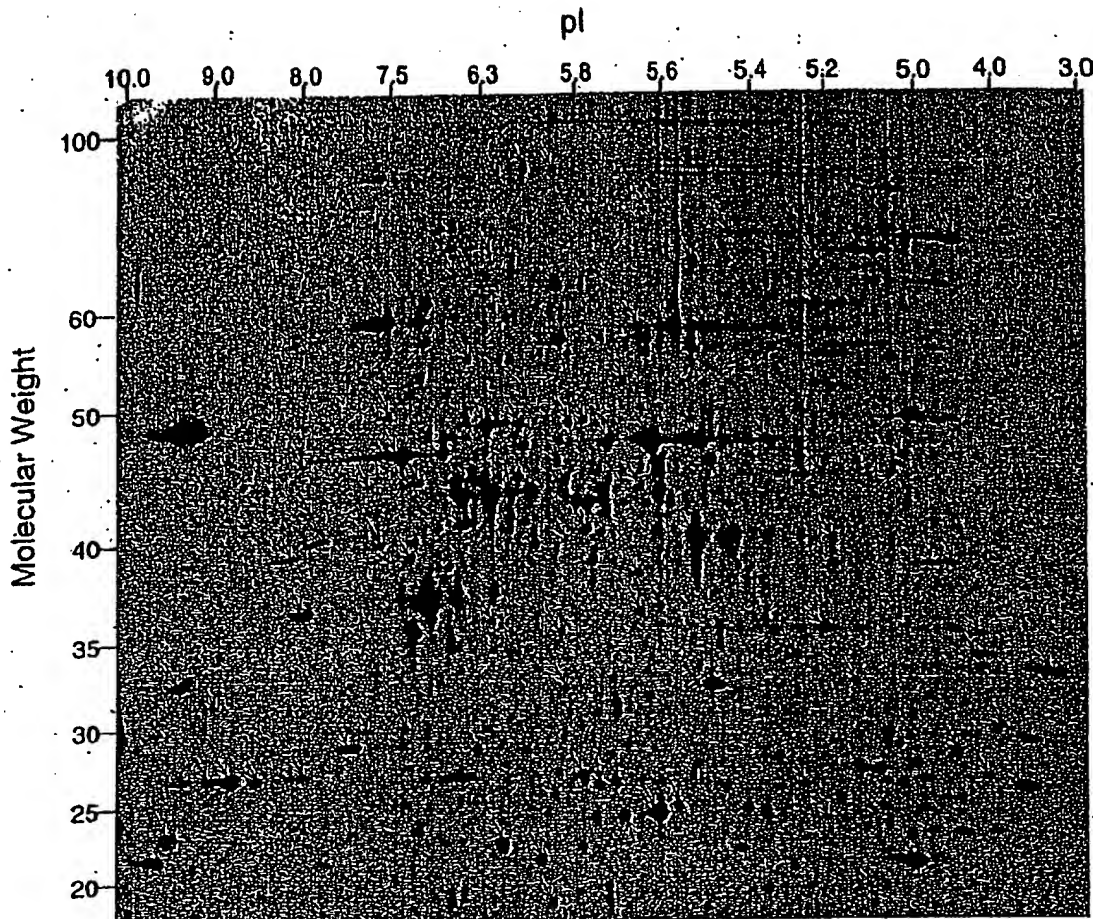


Figure 5. 2-DE separation of a lysate of yeast cells, with identified proteins highlighted. The first dimension of separation was an IPG from pH 3–10, and the second dimension was a 10%T SDS-PAGE gel. Proteins were visualized by silver staining. Further details of experimental procedures are included in S. P. Ouyi *et al.* (submitted).

calculated measure of the degree of redundancy of triplet DNA codons used to produce each amino acid in a particular gene sequence. It has been shown to be a useful indicator of the level of the protein product of a particular gene sequence present in a cell [46]. The general rule which applies is that the higher the value of the codon bias calculated for a gene, the more abundant the protein product of that gene becomes. The calculated codon bias values corresponding to the proteins identified in Fig. 5 are shown in Fig. 6b. Nearly all of the proteins identified (> 95%) have codon bias values of > 0.2, indicating they are highly abundant in cells. In contrast, codon bias values calculated for the entire yeast genome (Fig. 6a) show that the majority of proteins present in the proteome have a codon bias of < 0.2 and are thus of low abundance.

This finding is of considerable importance in our assessment of the current status of proteome analysis technology. It is clear that even using highly sensitive analytical techniques, we are only able to visualize and identify the

more abundant proteins. Since many important regulatory proteins are present only at low abundance, these would not be amenable to analysis using such techniques. This situation would be exacerbated in the analysis of proteomes containing many more proteins than the approximately 6000 gene products present in yeast cells [16]. In the analysis of, for example, the proteome of any human cells; there are potentially 50 000–100 000 gene products [47]. Inherent limitations on the amount of protein that can be loaded on 2-DE, and the number of components that can be resolved, indicate that only the most highly abundant fraction of the many gene products could be successfully analyzed. One approach that has been employed to circumvent these limitations is the use of very narrow range immobilized pH gradient strips for the first-dimension separation of 2-DE [48]. Since only those proteins which focus within the narrow range will enter the second dimension of separation, a much higher sample loading within the desired range is possible. This, in turn, can lead to the visualization and identification of less abundant proteins.

of a
for CE,
device,
pump,
The
servo
rannels
hanced
duced

imical
auto-
CID.
fy the
efore,
ber of
bilize
tra of
bases
l auto-
as can
y at a

es de-
spots
led by
tained
y high
d pro-
alysis
abun-
f the
is is a

c digest
microsa-
3, 290
tryptic
in LCQ
cted for
tified as
m car-
Repro-
2a.

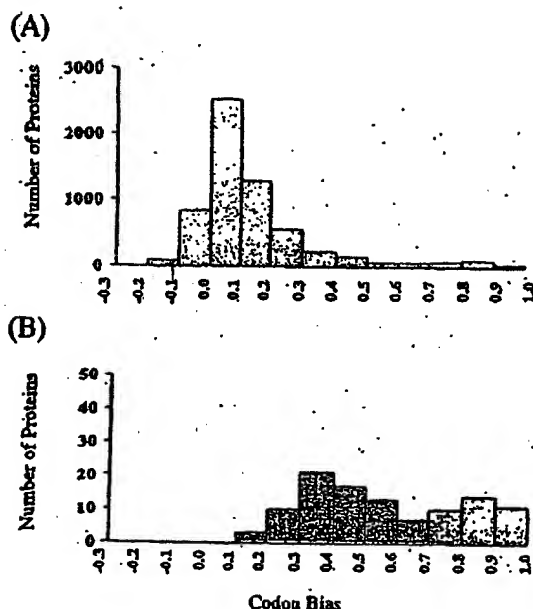


Figure 6. Calculated codon bias values for yeast proteins. (A) Distribution of calculated values for the entire yeast proteome. (B) Distribution of calculated values for the subset of 30 identified proteins also shown in Figs. 1 and 5. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

4 Utility of proteome analysis for biological research

For the success of proteomics as a mainstream approach to the analysis of biological systems it is essential to define how proteome analysis and biological research projects intersect. Without a clear plan for the implementation of proteome-type approaches into biological research projects the full impact of the technology can not be realized. The literature indicates that proteome analysis is used both as a database/data archive, and as a biological assay or biological research tool.

4.1 The proteome as a database

The use of proteomics as a database or data archive essentially entails an attempt to identify all the proteins in a cell or species and to annotate each protein with the known biological information that is relevant for each protein. The level of annotation can, of course, be extensive. The most common implementation of this idea is the separation of proteins by high resolution 2-DE, the identification of each detected protein spot and the annotation of the protein spots in a 2-DE gel database format. This approach is complicated by the fact that it is difficult to precisely define a proteome and to decide which proteome should be represented in the database. In contrast to the genome of a species, which is essentially static, the proteome is highly dynamic. Processes such as differentiation, cell activation and disease can all significantly change the proteome of a species. This is illustrated in Fig. 7. The figure shows two high-resolu-

tion 2-DE maps of proteins isolated from rat serum. Fig. 7A is from the serum of normal rats, while Fig. 7B is from the serum of rats in acute-phase serum after prior treatment with an inflammation-causing agent [49]. It is obvious that the protein patterns are significantly different in several areas, raising the question of exactly which proteome is being described.

Therefore, a comprehensive proteome database of a species or cell type needs to contain all of the parameters which describe the state and the type of the cells from which the proteins were extracted as well as the software tools to search the database with queries which reflect the dynamics of biological systems. A comprehensive proteome database should be capable of quantitatively describing the fate of each protein if specific systems and pathways are activated in the cell. Specifically, the quantity, the degree of modification, the subcellular location and the nature of molecules specifically interacting with a protein as well as the rate of change of these variables should be described. Using these admittedly stringent criteria, there is currently no complete proteome database. A number of such databases are, however, in the process of being constructed. The most advanced among them, in our opinion, are the yeast protein database YPD [50] (accessible at <http://www.ypd.com>) and the human 2D-PAGE databases of the Danish Centre for Human Genome Research [12] (accessible at <http://biobase.dk/cgi-bin/celis>). While neither can be considered complete as not all of the potential gene products are identified, both contain extensive annotation of supplemental information for many of the spots which are positively identified in reference samples.

4.2 The proteome as a biological assay

The use of proteome analysis as a biological assay or research tool represents an alternative approach to integrating biology with proteomics. To investigate the state of a system, samples are subjected to a specific process that allows the quantitative or qualitative measurement of some of the variables which describe the system. In typical biochemical assays one variable (e.g., enzyme activity) of a single component (e.g., a particular enzyme) is measured. Using proteomics as an assay, multiple variables (e.g., expression level, rate of synthesis, phosphorylation state, etc.) are measured concurrently on many (ideally all) of the proteins in a sample. The use of proteomics as an assay is a less far-reaching proposition than the construction of a comprehensive proteome database. It does, however, represent a pragmatic approach which can be adapted to investigate specific systems and pathways, as long as the interpretation of the results takes into account that with current technology not all of the variables which describe the system can be observed (see Section 3.4).

A common implementation of proteome analysis as a biological assay is when a 2-DE protein pattern generated from the analysis of an experimental sample is compared to an array of reference patterns representing different states of the system under investigation. The state of the experimental system at the time the sample was generated is therefore determined by the quantita-

rum.
: 7B
after
[49].
only
actly

spe-
sters
from
ware
flect
isive
lvely
tems
, the
locat-
ting
hese
edly
ome
r, in
uced
data-
and
entre
tp://
con-
pro-
ation
spots
des.

ty or
inte-
state
cess
ment
n. In
zyme
r en-
mul-
thesis,
ently
The
prop-
pro-
natic
ecific
on of
fnol-
stem

as a
gener-
le is
nting
. The
mple
ntita-

tive comparative analysis of hundreds to a few thousand proteins. Comparative analysis of the 2-DE patterns furthermore highlights quantitative and qualitative differences in the protein profiles which correlate with the state of the system. For this type of analysis it is not essential that all the proteins are identified or even visu-

alized, although the results become more informative as more proteins are compared. It is obvious, however, that the possibility to identify any protein deemed characteristic for a particular state dramatically enhances this approach by opening up new avenues for experimentation.

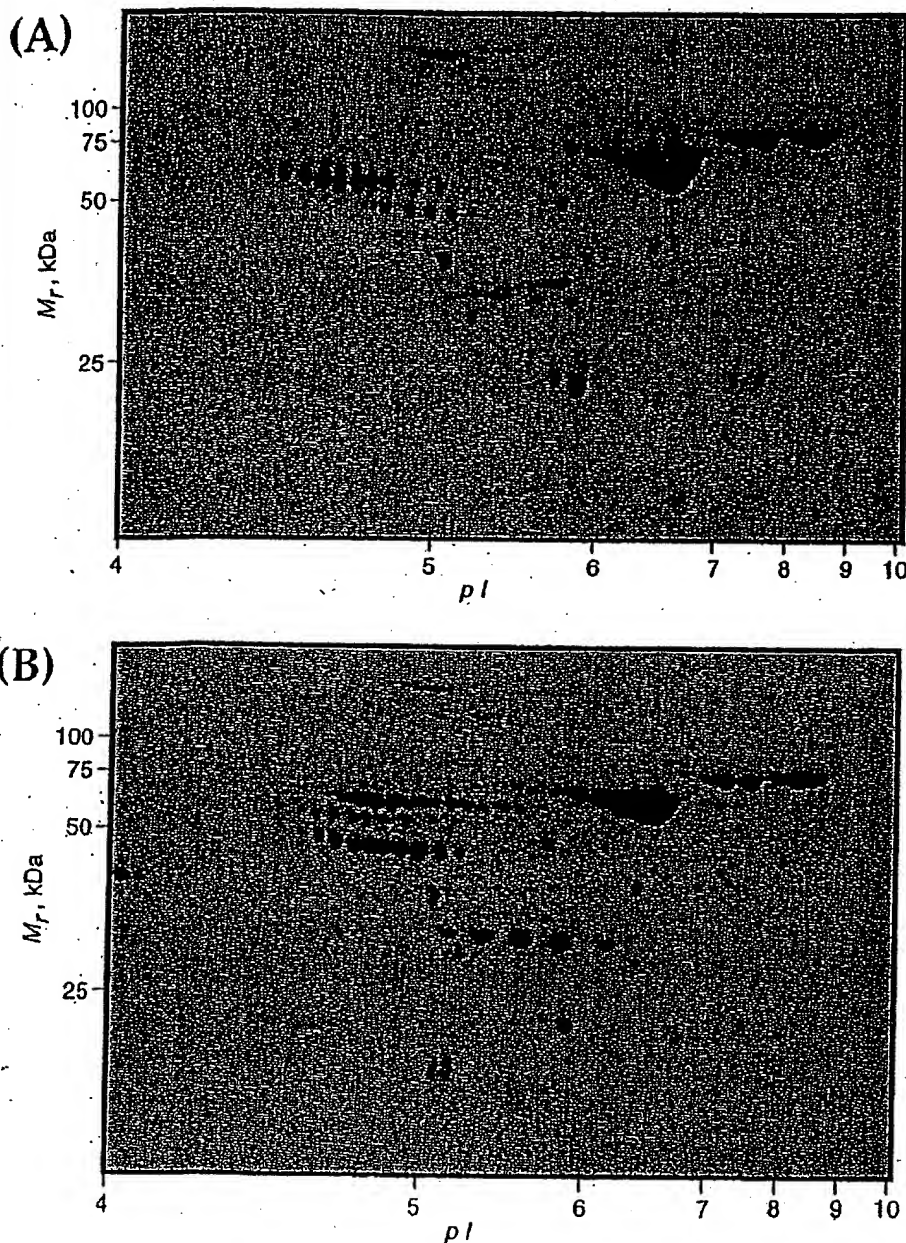


Figure 2. High resolution 2-DE map of proteins isolated from rat serum with or without prior exposure to an inflammation-causing agent. (A) normal rat serum, (B) acute-phase serum from rats which had previously been exposed to an inflammation-causing agent. The first dimension of separation is an IPG from pH 4–10, and the second dimension is a 7.5–17.5%T gradient SDS-PAGE gel. Proteins were visualized by staining with amido black. Further details of experimental procedures are included in [14, 49].

Proteome analysis as a biological assay has been successfully used in the field of toxicology, to characterize disease states or to study differential activation of cells. The approach is limited, of course, by the fact that only the visible protein spots are included in the assay; and it is well known that a substantial but far from complete fraction of cellular proteins are detected if a total cell lysate is separated by 2-DE. Proteins may not be detected in 2-DE gels because they are not abundant enough to be visualized by the detection method used, because they do not migrate within the boundaries (size, pI) resolved by the gel, because they are not soluble under the conditions used, or for other reasons.

A different way to use proteome analysis as a biological assay to define the state of a biological system is to take advantage of the wealth of information contained in 2-DE protein patterns. 2-DE is referred to as two-dimensional because of the electrophoretic mobility and the isoelectric points which define the position of each protein in a 2-DE pattern. In addition to the two dimensions used to generate the protein patterns, a number of additional data dimensions are contained in the protein patterns. Some of these dimensions such as protein expression level, phosphorylation state, subcellular location, association with other proteins, rate of synthesis or degradation indicate the activity state of a protein or a biological system. Comparative analysis of 2-DE protein patterns representing different states is therefore ideally suited for the detection, identification and analysis of suitable markers. Once again it must be emphasized that in this type of experiment only a fraction of the cellular proteins is analyzed. Since many regulatory proteins are of low abundance, this limitation is a concern, particularly in cases in which regulatory pathways are being investigated.

5 Concluding remarks

In this report we have addressed three main issues related to proteome analysis. First, we have discussed the rationale for studying proteomes. Second, we have assessed the technical feasibility of analyzing proteomes and described current proteome technology, and third, we have analyzed the utility of proteome analysis for biological research. It is apparent that proteome analysis is an essential tool in the analysis of biological systems. The multi-level control of protein synthesis and degradation in cells means that only the direct analysis of mature protein products can reveal their correct identities, their relevant state of modification and/or association and their amounts. Recently developed methods have enabled the identification of proteins at ever-increasing sensitivity levels and at a high level of automation of the analytical processes. A number of technical challenges, however, remain. While it is currently possible to identify essentially any protein spots that can be visualized by common staining methods, it is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions.

We have outlined the two principal ways proteome analysis is currently being used to intersect with biological research projects: the proteome as a database or data archive and proteome analysis as a biological assay. Both approaches have in common that at present they are conceptually and technically limited. Current proteome databases typically are limited to one cell type and one state of a cell and therefore do not account for the dynamics of biological systems. The use of proteome analysis as a biological assay can provide a wealth of information, but it is limited to the proteins detected and is therefore not truly proteome-wide. These limitations in proteomics are to a large extent a reflection of the fact that proteins in their fully processed form cannot easily be amplified and are therefore difficult to isolate in amounts sufficient for analysis or experimentation. The fact that to date no complete proteome has been described further attests to these difficulties. With continued rapid progress in protein analysis technology, however, we anticipate that the goal of complete proteome analysis will eventually become attainable.

We would like to acknowledge the funding for our work from the National Science-Foundation Science and Technology Center for Molecular Biotechnology and from the NIH. We thank Yvan Rochon and Bob Franza for providing the yeast gel shown and Elisabetta Gianazza for providing the rat serum gels shown.

Received April 21, 1998

6 References

- [1] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yao, J. X., Goolley, A. A., Hughes, O., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., *Bio/Technology* 1996, 14, 61-65.
- [2] Hodges, P. E., Payne, W. E., Garrels, J. L., *Nucleic Acids Res.* 1998, 26, 68-72.
- [3] O'Connor, C. D., Ferris, M., Fowler, R., Qi, S. Y., *Electrophoresis* 1997, 18, 1483-1490.
- [4] Cordwell, S. J., Basrael, D. J., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1335-1346.
- [5] Urquhart, B. L., Atsalos, T. E., Roach, D., Basrael, D. J., Bjellqvist, B., Britton, W. L., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1384-1392.
- [6] Wasinger, Y. C., Bjellqvist, B., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1373-1383.
- [7] Link, A. J., Hays, L. G., Carmack, R. B., Yates III, J. R., *Electrophoresis* 1997, 18, 1314-1334.
- [8] Suzuki, T., Ohara, O., *Electrophoresis* 1997, 18, 1252-1258.
- [9] VanBogelen, R. A., Abshire, K. Z., Moldover, B., Olson, E. R., Neldhardt, F. C., *Electrophoresis* 1997, 18, 1243-1251.
- [10] Guerreiro, N., Redmond, J. W., Rolfe, B. G., Djordjevic, M. A., *Mol. Plant Microbe Interact.* 1997, 10, 506-516.
- [11] Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Goolley, A. A., Hochstrasser, D. F., Williams, K. L., *Electrophoresis* 1997, 18, 491-497.
- [12] Cells, J., Gromov, P., Ostergaard M., Madsen, P., Honoré, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., Rasmussen, H. H., *FEBS Lett.* 1996, 394, 129-134.
- [13] Appel, R. D., Sanchez, J.-C., Baiocchi, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1232-1238.
- [14] Haynes, P., Miller, L., Aebersold, R., Oemmel, M., Eberlin, I., Lovati, R. M., Manzoni, C., Vignati, M., Gianazza, E., *Electrophoresis* 1998, 19, 1484-1492.

- [15] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-J., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, N. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghegan, N. S. M., Gish, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, C. O., Venter, J. C., *Science* 1995, 269, 496-512.
- [16] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. O., *Science* 1996, 274, 546.
- [17] Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Karlavaga, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Wathey, T., McDonald, L., Artlich, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., Venter, J. C., *Nature* 1997, 390, 580-586.
- [18] Liang, P., Pardee, A. B., *Science* 1992, 257, 967-971.
- [19] Lashkari, D. A., Dorris, J. L., McCusker, J. H., Namath, A. P., Genille, C., Hwang, S. Y., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, 94, 13057-13062.
- [20] Shalton, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, 6, 619-645.
- [21] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, 270, 484-487.
- [22] Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, B., Basrai, M. A., Bassett, D. B., Hieter, P., Vogelstein, B., Kinzler, K. W., *Cell* 1997, 88, 243-251.
- [23] Krishna, R. G., Wold, F., *Adv. Enzymol.* 1993, 67, 265-298.
- [24] Götz, A., Postel, W., Günther, S., *Electrophoresis* 1988, 9, 531-546.
- [25] Klose, J., Kobalz, U., *Electrophoresis* 1995, 16, 1034-1059.
- [26] Malsudala, P., *J. Biol. Chem.* 1987, 262, 10035-10038.
- [27] Aebersold, R. H., Teplow, D. B., Hood, L. E., Kent, S. B., *J. Biol. Chem.* 1986, 261, 4229-4238.
- [28] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, 203, 173-179.
- [29] Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B., *Proc. Natl. Acad. Sci. USA* 1987, 84, 6970-6974.
- [30] Hoopré, B., Leffers, H., Madsen, P., Celis, J. B., *Eur. J. Biochem.* 1993, 212, 421-430.
- [31] Mann, M., Wilm, M., *Anal. Chem.* 1994, 66, 4390-4399.
- [32] Eng, J., McCormack, A. L., Yates III, J. R., *J. Amer. Mass Spectrom.* 1994, 5, 976-989.
- [33] Yates III, J. R., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, 67, 1426-1436.
- [34] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, 68, 850-858.
- [35] Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., Aebersold, R., *Protein Sci.* 1991, 2, 1342-1351.
- [36] van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Biochem.* 1997, 247, 310-318.
- [37] Lui, M., Tempst, P., Erdjument-Bromage, H., *Anal. Biochem.* 1996, 241, 156-166.
- [38] Patterson, S. D., Aebersold, R. A., *Electrophoresis* 1995, 16, 1791-1814.
- [39] Ducret, A., Poy, Brunn, C., Bures, E. J., Marhaug, G., Husby, G. R. A., *Electrophoresis* 1996, 17, 866-876.
- [40] Haynes, P. A., Frigg, N., Aebersold, R., *Electrophoresis* 1998, 19, 939-945.
- [41] Figey, D., Van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, 68, 1822-1828.
- [42] Ducret, A., Van Oostveen, I., Bag, J. K., Yates III, J. R., Aebersold, R., *Protein Sci.* 1997, 7, 706-719.
- [43] Figey, D., Ducret, A., Yates III, J. R., Aebersold, R., *Nature Biotech.* 1996, 14, 1579-1583.
- [44] Figey, D., Aebersold, R., *Electrophoresis* 1997, 18, 360-368.
- [45] Figey, D., Ning, Y., Aebersold, R., *Anal. Chem.* 1997, 69, 3153-3160.
- [46] Garrett, J. L., McLaughlin, C. S., Warner, J. R., Fletcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., Payne, W. E., *Electrophoresis* 1997, 18, 1347-1360.
- [47] Schuler, G. D., Boguski, M. S., Stewart, B. A., Stein, L. D., Grapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birrea, B. B., Butler, A., Castle, A. B., Chianikukhal, N., Chu, A., Cleo, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.-B., Fang, N., Fitzames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, B., Hul, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGillivray, A., Mader, C., Maraskulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muschel, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, P., Quackenbush, J., Ranby, S., Roif, T., Rozen, S., Sanders, X., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehejia, A., Goodfellow, P. N., Houligatte, R., Hudson, J. R., Jr., Ide, S. E., Iorio, K. R., Leo, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., et al. *Science* 1996, 274, 540-546.
- [48] Sanchez, J.-C., Rouge, V., Pisteu, M., Raviez, F., Tonella, L., Moosmayer, M., Wilkins, M. R., Hochstrasser, D. P., *Electrophoresis* 1997, 18, 324-327.
- [49] Miller, I., Haynes, P., Gemelnet, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., Gianazza, E., *Electrophoresis* 1998, 19, 1493-1500.
- [50] Garrett, J. L., *Nucleic Acids Res.* 1996, 24, 46-49.

Gene-expression profiles predict survival of patients with lung adenocarcinoma

DAVID G. BEER¹, SHARON L.R. KARDIA², CHIANG-CHING HUANG³, THOMAS J. GIORDANO⁴, ALBERT M. LEVIN², DAVID E. MISEK⁵, LIN LIN¹, GUOAN CHEN¹, TAREK G. GHARIB¹, DAFYDD G. THOMAS⁴, MICHELLE L. LIZYNESS⁴, RORK KUICK⁵, SATORU HAYASAKA³, JEREMY M.G. TAYLOR³, MARK D. IANNETTONI¹, MARK B. ORRINGER¹ & SAMIR HANASH⁵

Departments of ¹Surgery, ²Epidemiology, ³Biostatistics, ⁴Pathology and ⁵Pediatrics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence should be addressed to D.G.B.; email: dgbeer@umich.edu.

Published online: 15 July 2002, doi:10.1038/nm733

Histopathology is insufficient to predict disease progression and clinical outcome in lung adenocarcinoma. Here we show that gene-expression profiles based on microarray analysis can be used to predict patient survival in early-stage lung adenocarcinomas. Genes most related to survival were identified with univariate Cox analysis. Using either two equivalent but independent training and testing sets, or 'leave-one-out' cross-validation analysis with all tumors, a risk index based on the top 50 genes identified low-risk and high-risk stage I lung adenocarcinomas, which differed significantly with respect to survival. This risk index was then validated using an independent sample of lung adenocarcinomas that predicted high- and low-risk groups. This index included genes not previously associated with survival. The identification of a set of genes that predict survival in early-stage lung adenocarcinoma allows delineation of a high-risk group that may benefit from adjuvant therapy.

Lung cancer remains the leading cause of cancer death in industrialized countries. Most patients with non-small cell lung cancer (NSCLC) present with advanced disease, and despite recent advances in multi-modality therapy, the overall 10-year survival rate remains a dismal 8–10%¹. However, a significant minority of patients (~25–30%) with NSCLC have stage I disease and receive surgical intervention alone. Although 35–50% of patients with stage I disease will relapse within 5 years^{2–4}, it is not currently possible to identify specific high-risk patients.

Adenocarcinoma is currently the predominant histological subtype of NSCLC (refs. 1,5,6). Although morphological assessment of lung carcinomas can roughly stratify patients, there is a need to identify patients at high risk for recurrent or metastatic disease. Preoperative variables that affect survival of patients with NSCLC have been identified^{7–10}. Tumor size, vascular invasion, poor differentiation, high tumor-proliferative index and several genetic alterations, including *K-ras* (refs. 11,12) and *p53* (refs. 10,13) mutations, have prognostic significance. Multiple independently assessed genes or gene products have also been investigated to better predict patient prognosis in lung cancer^{14–18}. Technologies that simultaneously analyze the expression of thousands of genes¹⁹ can be used to correlate gene-expression patterns with numerous clinical parameters—including patient outcome—to better predict tumor behavior in individual patients²⁰. Analyses of lung cancers using array technologies have identified subgroups of tumors that differ according to tumor type and histological subclasses and, to a lesser extent, survival among adenocarcinoma patients^{21,22}. Here we correlated gene-expression profiles with clinical outcome in a cohort of patients with lung adenocarcinoma and identified specific genes that

predict survival among patients with stage I disease. For further validation, we also show that the risk index predicted survival in an independent cohort of stage I lung adenocarcinomas.

Hierarchical profile clustering yields three tumor subsets

Using oligonucleotide arrays, we generated gene-expression profiles for 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, as well as 10 non-neoplastic lung samples. Selected sample replicates showed high correlation among coefficients and reliable reproducibility. We determined transcript abundance using a custom algorithm and the data set was trimmed of genes expressed at extremely low levels, that is, genes were excluded if the measure of their 75th percentile value was less than 100. Although potentially resulting in the loss of some information, trimming in this manner decreased the possibility that the clustering algorithm would be strongly influenced by genes with little or no expression in these samples. Hierarchical clustering with the resulting 4,966 genes yielded 3 clusters of tumors (Fig. 1). All 10 non-neoplastic samples clustered tightly together within Cluster 1 (data not shown). We examined the relationships between cluster and patient and tumor characteristics (Fig. 1 and Supplementary Figure A online). There were associations between cluster and stage ($P = 0.030$) and between cluster and differentiation ($P = 0.01$). Cluster 1 contained the greatest percentage (42.8%) of well differentiated tumors, followed by Cluster 2 (27%) and Cluster 3 (4.7%). Cluster 3 contained the highest percentage of both poorly differentiated (47.6%) and stage III tumors (42.8%), yet contained 3 (14.3%) moderately differentiated and 1 (5%) well differentiated stage I tumor. Notably, 11 stage I tumors were present in Cluster 3, sug-

gesting a common gene-expression profile for this subset of stage I and stage III tumors.

For patients with stage I and stage III tumors, the average ages were 68.1 and 64.5 years and the percentage of smokers was 88.9% and 89.5%, respectively. Marginally significant associations between cluster and smoking history were observed ($P = 0.06$). A significant relationship between histopathological classification and cluster was only discernable for bronchioloalveolar adenocarcinomas (BAs), which were only present in Clusters 1 and 2 ($P = 0.0055$) and comprised 35.7% and 12.3% of tumors for Clusters 1 and 2, respectively.

We examined the heterogeneity in gene-expression profiles based on the trimmed data set among normal lung samples and stage I and stage III adenocarcinomas by calculating correlation coefficients between all pairs of samples. In contrast to normal lung samples that displayed highly similar gene-expression profiles (median correlation, 0.9), both stage I and III lung tumors demonstrated much greater heterogeneity in their expression profiles with lower correlation coefficients (median values, 0.82 and 0.79, respectively).

Northern-blot and immunohistochemistry analyses

Of the 4,966 genes examined, 967 differed significantly between stage I and III adenocarcinomas, a number in excess of that expected by chance alone (248 at alpha level (α) = 0.05). Three genes were arbitrarily selected to verify the microarray expression data. The mRNA from 20 of the normal lung and tumor samples was examined by northern-blot hybridization with probes for insulin-like growth factor-binding protein 3 (*IGFBP3*), cystatin C

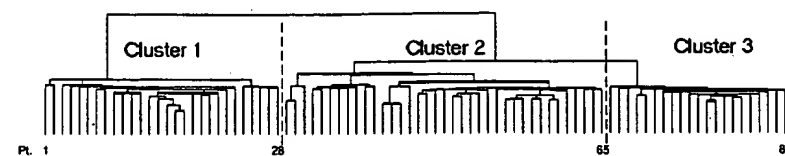


Fig. 1 Unsupervised classification analysis of lung adenocarcinomas. 3 classes of tumors identified by agglomerative hierarchical clustering of gene-expression profiles using the 4,966 expressed genes. Patient and histopathological information for each lung adenocarcinoma case by cluster designation and methods for *K-ras* 12/13th-codon mutational status and nuclear p53 protein accumulation are provided (Supplementary Figure A online). TN classification denotes information regarding patient tumor size and nodal involvement. Associations between cluster membership and patient or histopathological variables are indicated at significance level ($P \leq 0.05$).

and lactate dehydrogenase A (*LDH-A*) (Fig. 2a). Two gene probes not represented on the microarrays were used as controls, including histone H4, a potential index of overall cell proliferation, and 28S ribosomal RNA, a control for sample loading and transfer. The relative amounts of *IGFBP3*, cystatin C and *LDH-A* mRNA strongly correlated with microarray-based measurements (Fig. 2b). In both assays, *IGFBP3* and *LDH-A* mRNA levels increased from stage I to stage III adenocarcinomas and were higher than those in normal lung. Cystatin C mRNA levels were more variable but relatively greater in normal lung than tumors. These results suggest that the oligonucleotide microarrays provided reliable measures of gene expression. The tumors showed slightly greater histone H4 expression than the normal lung, likely reflecting increased proliferation of tumor cells.

Immunohistochemistry was performed for *IGFBP3*, cystatin C and HSP-70 to determine whether mRNA overexpression was reflected by an increase of their corresponding proteins in tumors.

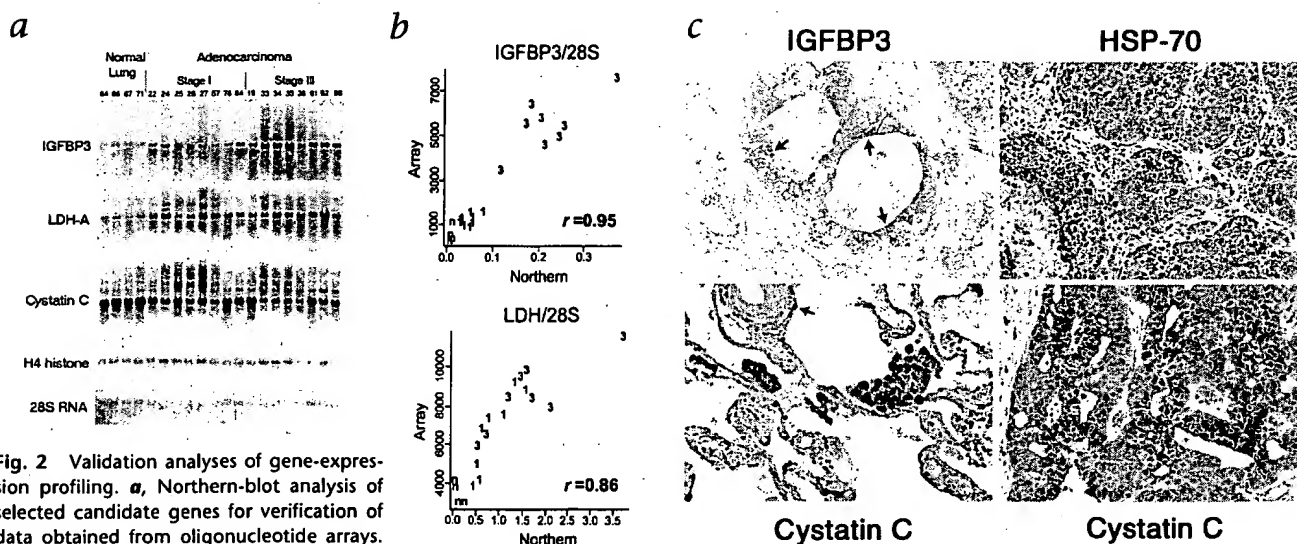


Fig. 2 Validation analyses of gene-expression profiling. **a**, Northern-blot analysis of selected candidate genes for verification of data obtained from oligonucleotide arrays. The same sample RNA for the 4 uninvolved lung, 8 stage I and 8 stage III tumors was used for the northern-blot and oligonucleotide array analyses. **b**, Correlation analysis of quantitative data obtained from oligonucleotide arrays and northern blots measured by integrated phosphorimager-based signals for the *IGFBP3* and *LDH-A* genes. The ratio of *IGFBP3*, cystatin C and *LDH-A* mRNA to 28S rRNA was determined. The relative values for each gene from each sample are shown. n, non-neoplastic normal lung; 1, stage I tumors; 3, stage III tumors. **c**, Immunohistochemical analysis of *IGFBP3*, HSP-70 and cystatin C in lung and lung adenocarcinomas. Cytoplasmic *IGFBP3* immunoreactivity in a neoplastic gland (tumor L22)

with prominent apical staining (blue reactant staining, arrow, upper left). Diffuse cytoplasmic HSP-70 immunoreactivity (tumor L27), yet stromal elements show no reactivity (upper right). Normal lung parenchyma (lower left) shows cytoplasmic cystatin C immunoreactivity in alveolar pneumocytes (arrow) and intra-alveolar macrophages but tumor (L90) shows diffuse cytoplasmic cystatin C immunoreactivity with prominent apical staining (lower right). Magnification, $\times 200$

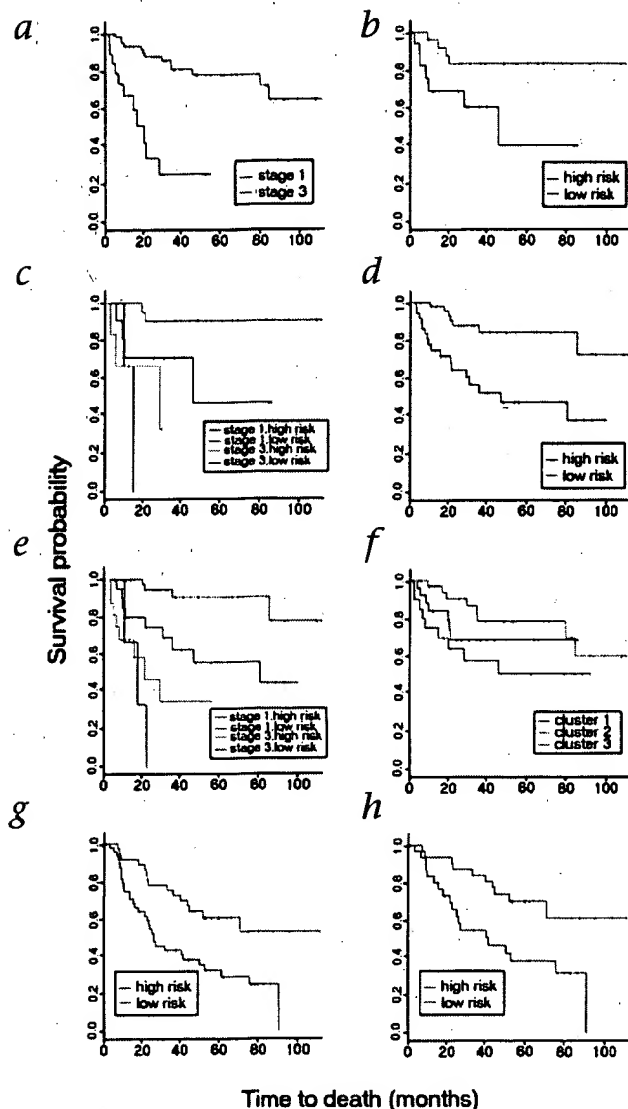


Fig. 3 Gene-expression profiles and patient survival. **a**, Relationship between tumor stage and patient survival (stage 1 and stage 3 differ significantly, $P < 0.0001$). **b**, Relationship between the survival in the 43 test samples and their risk assignments based on the 50-gene risk index estimated in the 43 training samples. The high- and low-risk groups differ significantly ($P = 0.024$). **c**, Relationship between patient survival and the risk assignments in test samples (in **b**) conditional for tumor stage. The high- and low-risk stage I groups differ significantly ($P = 0.028$), whereas stage III low- and high-risk groups did not ($P = 0.634$). **d**, Relationship between survival in the test cases and their risk assignments based on the 86 'leave-one-out' cross-validation of the 50-gene risk index. The high- and low-risk groups differ significantly ($P = 0.0006$). **e**, Relationship between test case's risk assignment and survival (in **d**) conditional on tumor stage. The high- and low-risk stage I lung adenocarcinoma groups differ significantly from each other ($P = 0.003$), whereas low- and high-risk stage III tumors do not. **f**, Relationship between tumor class identified by hierarchical clustering and patient survival. Survival for patients in Cluster 3 differed relative to the tumors in Cluster 2 ($P = 0.037$) and approached significance for Cluster 1 and 2 combined ($P = 0.06$). **g**, Analysis of the Michigan-based risk index using top cross-validated survival genes identify a low- and high-risk group in an independent cohort of 84 Massachusetts-based lung adenocarcinomas that are significantly different ($P = 0.003$). **h**, Among the 62 stage I lung adenocarcinomas in the Massachusetts sample, the high- and low-risk groups differed significantly ($P = 0.006$).

After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set ($P = 0.024$) (Fig. 3b and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3c), low- and high-risk groups among stage I tumors were found to differ ($P = 0.028$) in their survival.

Identification of a robust set of survival genes

Although predictive of patient survival, a single training-testing set may not provide the most robust set of genes due to random sampling issues. Therefore, a 'leave-one-out' cross-validation approach was used to identify genes associated with survival from all 86-tumor samples. We first developed a 50-gene risk index in each training set, and then applied the risk index to the test case held out from the full set of tumors and assigned the held out tumor to the high- or low-risk groups (Fig. 3d). The high and low-risk subgroups determined in the test cases differed significantly in their overall survival ($P = 0.0006$). Among the larger group of stage I lung adenocarcinomas, the low-risk ($n = 46$) and high-risk ($n = 21$) groups had markedly different survival ($P = 0.003$) (Fig. 3e). Table 1 lists selected examples of the cumulative top 100 genes derived from this cross-validation procedure (complete list in Supplementary Table A online).

It was also noted that many of the stage I patients in the high-risk subgroup (Fig. 3e) were present in Cluster 3 (Fig. 1). Kaplan-Meier analysis (Fig. 3f) demonstrated a significantly worse survival ($P = 0.037$) for patients in Cluster 3 relative to patients in Cluster 2 and approaching significance for Cluster 1 and 2 combined ($P = 0.06$). This further indicates the important relationship between gene-expression profiles and patient survival, independent of disease stage.

Consistent with previous analyses of lung adenocarcinomas²³, 40% of stage I and 57.8% of stage III tumors had 12th or 13th codon *K-ras* gene mutations. Those patients with tumors containing *K-ras* mutations showed a trend of poorer survival, but

Immunoreactivity for both *IGFBP-3* and *HSP-70* (Fig. 2c) was detected in the cytoplasm of the adenocarcinomas, with little detectable reactivity in the stromal or inflammatory cells. Cystatin C was detected in alveolar pneumocytes and intra-alveolar macrophages in non-neoplastic lung parenchyma and also consistently in the cytoplasm of neoplastic cells.

Gene-expression profiles predict survival

As expected, Kaplan-Meier survival curves (Fig. 3a) and log-rank tests indicated poorer survival among stage III compared with stage I adenocarcinomas ($P = < 0.0001$). Two statistical approaches were used to determine whether gene-expression profiles could predict survival using the data set of 4,966 genes. In one approach, equal numbers of randomly assigned stage I and stage III tumors constituted training ($n = 43$) and testing ($n = 43$) sets. In the training set, the top 10, 20, 50 or 75 genes were used to create risk indices that were evaluated for their association with survival using the 50th, 60th or 70th percentile cutoff points to categorize patients into high or low groups. The results were similar across cutoff points but the 50-gene risk index had the best overall association with survival in the training set.

Table 1 Selected examples of the top 100 genes from cross-validation

Gene name	P (normal versus tumor t-test)	% Change in tumor	P (stage I versus stage III t-test)	% Change in stage III	Coefficient β	Unigene comment
CASP4	0.56	-6%	0.02	57%	0.0022	Apoptosis-related Caspase 4, apoptosis- related cysteine protease
P63	9.73E-04	37%	0.03	43%	0.0010	Transmembrane protein (63 kD), endoplasmic reticulum/ Golgi intermediate compartment
KRT7	8.02E-08	126%	0.11	55%	0.0003	Cell adhesion and structure Keratin 7
LAMB1	0.14	-20%	0.01	60%	0.0027	Laminin, β 1
BMP2	0.54	-21%	0.27	47%	0.0044	Cell cycle and growth regulators Bone morphogenetic protein 2
CDC6	1.31E-05	1070%	0.05	148%	0.0124	CDC6 (cell division cycle 6, <i>Saccharomyces cerevisiae</i> homolog)
S100P	2.10E-08	1572%	0.19	77%	0.0001	S100 calcium-binding protein P
SERPINE1	2.89E-03	72%	0.25	30%	0.0008	Serine (or cysteine) proteinase inhibitor, clade E (nexin)
STX1A	8.65E-08	54%	0.07	26%	0.0031	Syntaxin 1A (brain)
ADM	0.05	39%	0.04	117%	0.0016	Cell signaling adrenomedullin
AKAP 12	8.53E-03	-47%	0.05	214%	0.0010	A kinase (PRKA) anchor protein (gravin) 12
ARHE	0.06	-39%	0.05	87%	0.0092	ras homolog gene family, member E
GRB7	2.02E-03	38%	0.63	15%	0.0030	Growth factor receptor-bound protein 7
VEGF	6.50E-08	174%	0.02	85%	0.0013	Vascular endothelial growth factor
WNT10B	0.05	31%	0.48	20%	0.0022	Wingless-type MMTV integration site family, member 10B
HSPA8	0.36	8%	9.01E-04	51%	0.0008	Chaperones Heat-shock 70 kD protein 8
ERBB2	0.04	92%	0.37	120%	0.0013	Receptors v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
FXD3	0.10	111%	0.31	73%	0.0046	FXD domain-containing ion transport regulator 3
SLC20A1	1.34E-03	58%	0.02	66%	0.0021	Solute carrier family 20 (phosphate transporter), member 1
CSTB	1.57E-04	50%	0.15	34%	0.0001	Enzymes, cellular metabolism Cystatin B (stefin B)
CTSL	0.48	-10%	0.03	67%	0.0007	Cathepsin L
CYP24	3.16E-06	N/A	0.97	2%	0.0008	Cytochrome P450, subfamily XXIV (vitamin D 24-hydroxylase)
FUT3	1.07E-07	114%	0.97	-1%	0.0033	Fucosyltransferase 3 (galactoside 3(4)-L- fucosyltransferase, Lewis blood group included)
MLN64	0.20	32%	0.42	80%	0.0007	Steroidogenic acute regulatory protein related
PDE7A	0.12	33%	0.01	-35%	-0.0187	Phosphodiesterase 7A
PLGL	0.04	-68%	0.35	-170%	-0.0011	Plasminogen-like
SLC1A6	0.07	-32%	0.12	86%	0.0069	Solute carrier family 1 (high-affinity aspartate/ glutamate transporter), member 6
COPEB	0.10	-33%	0.26	25%	0.0016	Transcription and translation Core promoter element binding protein
CRK	0.10	32%	0.03	48%	0.0098	v-crK avian sarcoma virus CT10 oncogene homolog
RELA	0.26	-7%	0.01	20%	0.0034	v-rel avian reticuloendotheliosis viral oncogene homolog A
KIAA0005	2.21E-04	40%	0.02	45%	0.0010	Unknown function KIAA0005 gene product
MGB1	0.27	125%	0.33	459%	0.0018	Mammaglobin 1

Bolded genes were also significant for survival in 43 tumor training set (Fig. 3b).

Table 1 Selected examples of the cumulative top 100 genes identified using training-testing, cross-validation of all 86 lung tumor samples. The percent change, as well as the direction, for the average values of the 10 non-neoplastic lung to all tumors, and for the 67 stage I to the 19 stage III tumors are shown. A positive coefficient β value is indicative of a relationship of gene expression to a

poorer patient outcome. The genes are listed in potential functional categories. Genes that were also present in the top 50 survival genes using the 43-tumor training set (Fig. 3b) are indicated in bold type. Complete listing of the gene probe sets and annotated gene and unigene identifiers can be found in the Supplementary Methods.

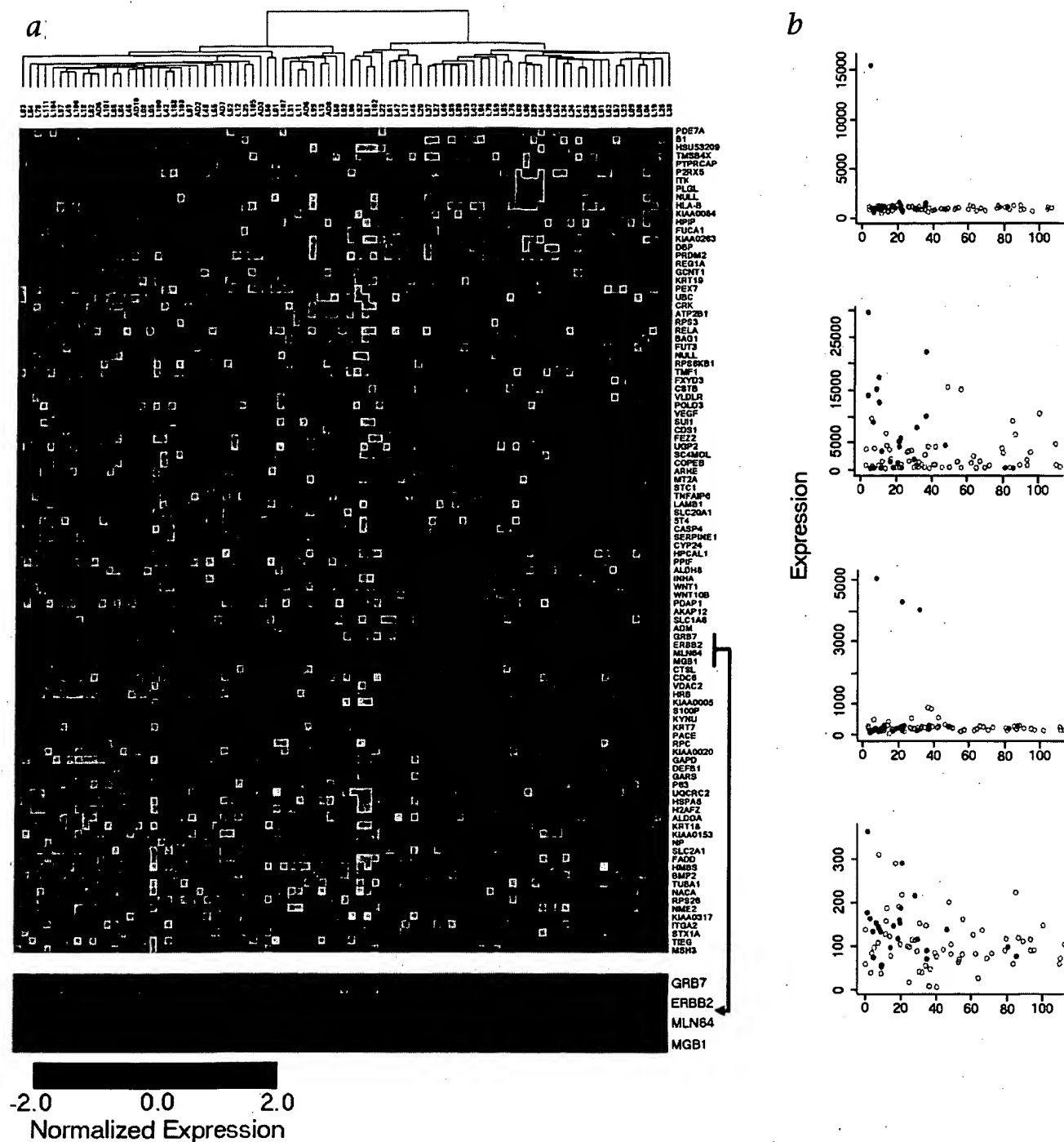


Fig. 4 Gene expression patterns of top survival genes **a**, Gene-expression patterns determined using agglomerative hierarchical clustering of the 86 lung adenocarcinomas against the 100 survival-related genes (Table 1) identified by the training-testing, cross-validation analysis. Substantially elevated (red) or decreased (green) expression of the genes is observed in individual tumors. Some tumors (black arrow and expanded area) show extremely elevated expression of specific genes. **b**, An outlier gene-expression pattern (>5 times the interquartile range among all samples) is observed for the *erbB2* and *Reg1A* genes (top left and right, respectively). The *S100P* and *crk* genes (bottom left and right, respectively) show a graded pattern of expression related to patient survival. ○, alive; ●, dead (also in **c**). **c**, The number of outliers per person identified in the top 100 genes plotted by survival distribution.

this difference did not reach statistical significance among all patients ($P = 0.25$), between patients within tumor clusters ($P = 0.41$) or when analyzed separately among stage I ($P = 0.22$) and stage III ($P = 0.53$) patients. Nuclear accumulation of p53 was detected in 17.9% stage I and in 22.2% stage III tumors. No significant relationship was observed for p53 staining and patient survival, cluster or tumor stage.

Confirmation using an independent set of adenocarcinomas

The robustness of our 50-gene risk index in predicting survival in lung adenocarcinomas was tested using oligonucleotide gene-expression data obtained from a completely independent (Massachusetts-based) sample of 84 lung adenocarcinomas (62 stage I, 14 stage II and 8 stage III; ref. 21, and dataset A at www.genome.wi.mit.edu/MPR/lung). To ensure equivalent power for testing and comparability of samples, the criteria for including tumors in the analysis were 40% or greater tumor cellularity, no mixed histology (that is, adenosquamous) and patient survival information. To obtain comparative gene-expression measures between the two data sets, gene sequences present on the U95A and HuGeneFL array were examined, and expression data for our top 50 cross-validation genes for all 84 Massachusetts samples were obtained and processed²⁴ (see also Supplementary Methods online). When we examined the risk assignment of these 84 samples, employing the identical cutoff point used for the 86 Michigan-based lung samples, we observed low- and high-risk groups (Fig. 3g; $P = 0.003$). Notably, among the 62 stage I tumors, high- and low-risk groups were observed that differed significantly ($P = 0.006$) in their survival (Fig. 3h).

Survival genes had graded and outlier expression patterns

A statistical and graphical analysis of the 100 survival-related

genes (Table 1) clustered against all 86 tumors revealed individual tumors with substantially elevated expression in both a limited and larger number of genes (Fig. 4a). Among these genes, we observed two distinct patterns of expression related to patient survival. One pattern, designated 'outlier', included genes showing substantially elevated expression (greater than five times the interquartile range among all samples), whereas the other pattern, designated 'graded', was characterized by continuously distributed expression with patient survival (Fig. 4b). The *erbB2* and *Reg1A* genes are examples of outlier expression patterns and *S100P* and *crk* genes of graded patterns. The number of outliers per person in the top 100 genes was identified and plotted according to survival times and events (Fig. 4c). Both stage I and stage III lung adenocarcinomas showed outlier gene patterns and 10 tumors contained 3 or more outlier genes.

Because gene amplification may result in increased gene expression, the nine genes with outlier expression patterns (*erbB2*, *SLC1A6*, *Wnt 1*, *MGB1*, *Reg1A*, *AKAP12*, *PACE*, *CYP24*, *KYNU*) and one gene with a graded expression pattern (*KRT18*) were examined using quantitative genomic PCR to evaluate genomic copy number (Fig. 5a). Gene amplification of *erbB2* (17q12) was detected in tumor L94, which had the highest *erbB2* mRNA expression (Fig. 4a). Gene amplification was not detected for any of the other seven tested genes in tumor L94, as well as in other tumors. The two genes most frequently demonstrating the outlier pattern in these lung adenocarcinomas were *KYNU* and *CYP24*, and were present in 10 and 9 tumors, respectively. *CYP24* has been described as a gene amplified and overexpressed in breast cancer²⁵, and these results indicate elevated expression in lung adenocarcinoma.

To determine whether the graded or outlier gene-expression patterns also occur at the protein-expression level, 10 of the 100

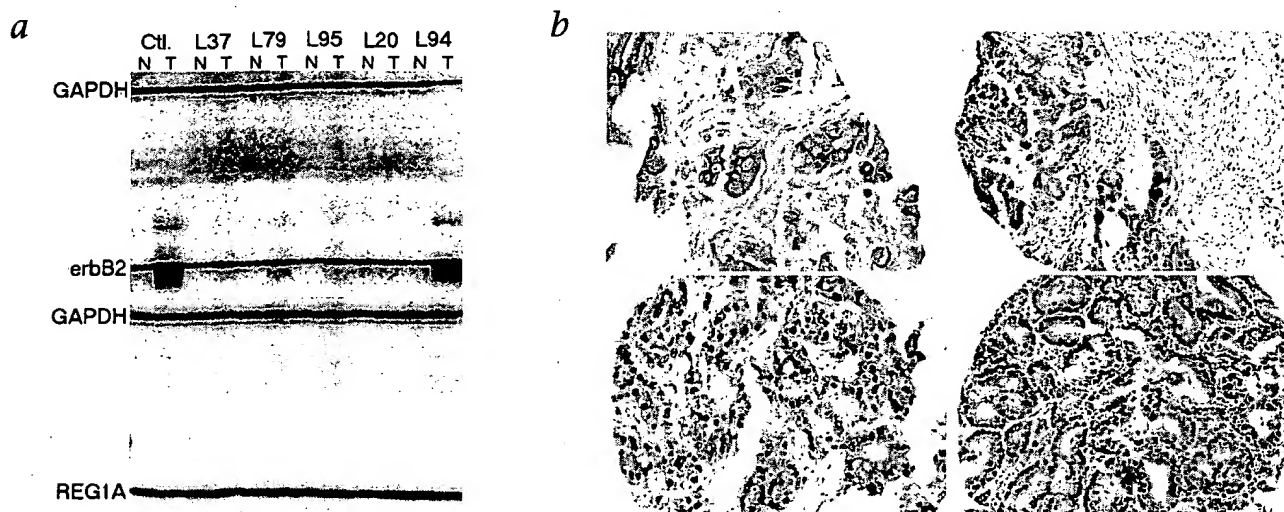


Fig. 5 Gene amplification and protein expression of survival-related genes. **a**, Analysis of potential gene amplification for 9 genes showing outlier expression patterns in the lung tumors (*erbB2*, *SLC1A6*, *Wnt 1*, *MGB1*, *Reg1A*, *AKAP12*, *PACE*, *CYP24* and *KYNU*) and examined using quantitative genomic PCR. A gene showing graded expression pattern (*KRT18*), and one gene (*PACE4*) with a similar chromosome location as *PACE*, were used as controls. Only *erbB2* and *Reg1A* are shown. An esophageal adenocarcinoma with known high-level genomic amplification of *erbB2* was used as a positive control and normal esophagus DNA was used as a negative control (Ct). PCR fragment sizes were 343 bp for *GAPDH*, 166 bp for *erbB2* and 126 bp for

Reg1A. DNA is from normal lung (N) and tumor (T) from each patient (for example L37). **b**, Immunohistochemical analysis of survival related genes with lung adenocarcinoma microarrays using the tumors from this study. The transmembrane *erbB2* protein (top left) expression is substantially increased in tumor L94 containing the amplified *erbB2* gene (Fig. 4a and b). Expression of *VEGF* (top right) and *S100P* (bottom left) was located within the neoplastic cells and the pattern of immunoreactivity was consistent with the graded expression pattern demonstrated by their mRNA profiles. Expression of the oncogene *crk* (bottom right) was abundantly expressed in neoplastic lung cells. Magnification, $\times 400$ (*erbB2*); $\times 200$ (*VEGF*, *S100P* and *crk*).

top survival genes (Table 1) for which specific antibodies were available were chosen for immunohistochemical analysis using lung-tumor arrays from this study (Fig. 5b). Expression of membrane *erbB2* protein was substantially increased in the *erbB2*-amplified tumor L94 and very low levels of expression were present in other tumors, consistent with mRNA-expression measurements (Fig. 4a and b). CDC6 protein expression was also substantially higher in tumor L94, consistent with mRNA levels (data not shown). Expression of vascular endothelial growth factor (VEGF) and S100P (Fig. 5b), as well as cytokeratin 18 (KRT18), cytokeratin 7 (KRT7) and fas-associated death domain (FADD) protein (data not shown), was located within the lung tumor cells and consistent with the graded expression pattern of the mRNA profiles. The oncogene *crk* showed both graded mRNA as well as a graded protein-expression pattern with survival, and was abundantly expressed in the tumor cells (Fig. 5b). These results indicate that many survival-associated genes are expressed at the protein level and demonstrate similar mRNA and protein-expression patterns.

Discussion

We used several approaches for the analysis of gene-expression data related to clinicopathological variables and patient survival. One approach, hierarchical clustering, was used to examine similarities among lung adenocarcinomas in their patterns of gene expression. Previous studies of lung tumors^{21,22} have also used this method to describe subclasses of lung tumors. Here, we found three clusters that showed significant differences with respect to tumor stage and tumor differentiation. This suggests, as expected, that tumors with similar histological features of differentiation demonstrate similarities in gene expression. This feature also partly underlies the observed statistical association of tumor stage and cluster, as many of the higher-stage tumors, often poorly differentiated and previously associated with a reduced survival^{9,10}, were located in Cluster 3. Although this cluster contained the highest percentage of stage III tumors, it also contained a nearly equal mixture of stage I and stage III tumors and not all tumors were poorly differentiated. This indicates that a subset of stage I lung adenocarcinomas share gene-expression profiles with higher-stage tumors. Notably, 10 of the 11 stage I tumors found in Cluster 3 were the high-risk stage I tumors identified using the risk index in the 'leave-one-out' cross-validation.

In contrast to previous analyses of lung adenocarcinomas^{21,22}, we validated the expression data from the arrays. The strong correlation of northern-blot analysis and oligonucleotide-array data for gene expression in the same samples (Fig. 2b) indicates that these studies provide robust gene-expression estimates. Immunohistochemistry using the same tumor samples in tissue arrays demonstrates protein expression within the lung tumor cells. Together, these studies indicate that many of the genes identified using gene-expression profiles are likely relevant to lung adenocarcinoma. For example, *IGFBP3* gene expression is increased in lung adenocarcinomas (Fig. 2c). *IGFBP3* protein modulates the autocrine or paracrine effects of insulin-like growth factors, elevated *IGFBP3* expression is observed in colon cancer²⁶, and increased serum *IGFBP3* is associated with progression in breast cancer²⁷. Heat-shock protein 70 (HSP-70) is increased in lung adenocarcinomas of smokers²⁸ and is associated with increased metastatic potential in breast cancer²⁹. Increased serum lactate dehydrogenase is correlated with tumor stage and tumor burden³⁰, and cystatin C, a cysteine protease inhibitor ex-

pressed in human lung cancers³¹, is prognostic in some cancers³². The decreased expression of this protease inhibitor may affect the invasive properties of the tumor cell.

The cross-validation analytical strategy we used is particularly informative for these types of gene-expression analyses for disease outcome^{33,34}, and identification of cross-validated genes with a larger tumor cohort may help refine this risk index for use in a clinical setting. The gene-expression data also provide opportunities to observe overarching patterns that advance our understanding of associations between genes and disease. For example, the top 100 survival genes include those involved in signaling, cell cycle and growth, transcription, translation and metabolism. Expression of many of these genes is likely a function of increased proliferation and metabolism in the more aggressive tumors. Some genes, such as *erbB2* and *Reg1A* (Fig. 4a and b), were highly overexpressed in a few patients having poor survival. In one tumor, the *erbB2* gene was amplified (Fig. 5a), demonstrating that genomic changes may underlie the overexpression of a subset of these outlier genes. Immunohistochemistry confirmed protein overexpression in this patient's tumor (Fig. 5b). Notably, seven of the eight outlier genes were not amplified, indicating that other mechanisms underlie the increased mRNA expression of these survival-related genes.

Most genes showed a graded relationship between expression and patient survival. Genes such as that encoding VEGF, known to be strongly associated with survival in lung cancer^{35,36} were identified as related to patient survival in our study. *VEGF* demonstrated a graded expression pattern, as did the *S100P* and *crk* oncogene (Fig. 5b). *S100P* is a calcium-regulated protein not previously reported in lung cancer. The *crk* gene, the cellular homolog of the *v-crak* oncogene, is a member of a family of adaptor proteins involved in signal transduction and interacts directly with c-jun N-terminal kinase 1 (JNK1)³⁷. Although *crk* has not been shown to have a role in lung cancer, its role in the MAP-kinase pathway, which leads to activation of matrix metalloproteinase secretion and cell invasion³⁸, indicates potential involvement in the tumor cell invasion or metastasis of some lung adenocarcinomas. Among the many genes identified in this study, like *crk*, that may be causally involved in lung cancer progression (Table 1), some were related to survival in many patients, and others in only smaller subsets of patients. This result is consistent with the complex molecular architecture of tumors in general, the heterogeneity of lung adenocarcinomas in particular and the multiple mechanisms underlying tumor-cell survival, invasion and metastasis³⁹.

Our results demonstrate that a gene-expression risk profile—based on the genes most associated with patient survival—can distinguish stage I lung adenocarcinomas and differentiate prognoses. The particular genes that define the clusters, or are associated with survival, likely reflect the characteristics of the particular tumors included in the analysis. Current therapy for patients with stage I disease usually consists of surgical resection without adjuvant treatment⁴³. Clearly, the identification of a high-risk group among patients with stage I disease would lead to consideration of additional therapeutic intervention for this group, possibly leading to improved survival of these patients.

Methods

Patient population. Sequential patients seen at the University of Michigan Hospital between May 1994 and July 2000 for stage I or stage III lung adenocarcinoma were evaluated for this study. Consent was received and the project was approved by the local Institutional Review Board. Primary tumors and adjacent non-neoplastic lung tissue were obtained at the time of

surgery. Peripheral portions of resected lung carcinomas were sectioned, evaluated by a study pathologist and compared with routine H&E sections of the same tumors, and utilized for mRNA isolation. Regions chosen for analysis contained a tumor cellularity greater than 70%, no mixed histology, potential metastatic origin, extensive lymphocytic infiltration or fibrosis. Tumors were histopathologically divided into two categories based on their growth pattern: bronchial-derived, if they exhibited invasive features with architectural destruction, and bronchioloalveolar, if they exhibited preservation of the lung architecture. All stage I patients received only surgical resection with intra-thoracic nodal sampling and no other treatments. Stage III patients received surgical resection plus chemotherapy and radiotherapy.

Gene-expression profiling and K-ras mutation analysis. RNA isolation, cRNA synthesis and gene-expression profiling were performed as described²⁴. Details of gene annotation and K-ras mutation analysis are provided in supplementary information.

Northern-blot analysis. Total cellular RNA (10 µg) was separated in 1.2% agarose-formaldehyde gels and vacuum-transferred to Gene Screen Plus (NEN Life Science Products, Boston, Massachusetts). Hybridization conditions and probe labeling were as described⁴⁰. Individual sequence-validated cDNA image clones for human IGFBP3 (clone 1407750), LDH-A (clone 2420241), cystatin C (CTS3; clone 949938) were from Research Genetics (Huntsville, Alabama). The human histone H4 cDNA and the 28S ribosomal RNA 26-mer oligonucleotide probe were prepared and labeled as described⁴⁰.

Gene-amplification analysis. 11 genes were selected for the analysis of genomic alterations. Primers were designed using PrimerSelect 4.05 Windows 32 software (DNASTAR, Madison, Wisconsin), avoiding pseudogenes or potential homologous regions. Forward and reverse primers for the genes are provided (Supplementary Methods online). Quantitative genomic-PCR was then applied and analyzed as described⁴¹.

Immunohistochemical staining. The H&E-stained slides of all primary lung tumors were used to identify the most representative regions of each tumor and a tissue microarray (TMA) block was constructed as described⁴². Immunohistochemistry (IHC) was performed using both routine and sections from the TMA block as described²⁴. Detailed methods and the concentrations used for all antibodies are provided in the Supplementary Methods.

Statistical methods. t-tests were used to identify differences in mean gene-expression levels between comparison groups. Agglomerative hierarchical clustering⁴³ was applied using the average linkage method to investigate whether there was evidence for natural groupings of tumor samples based on correlations between gene-expression profiles. To investigate the robustness of the clustering inference, gene-expression values were perturbed by adding random Gaussian error of magnitude obtained from a duplicate sample to each data point and then reclustered to determine concordance in the tumor's class membership. Pearson, χ^2 and Fisher's exact tests were used to assess whether cluster membership was associated with physical and genetic characteristics of the tumors.

To determine whether gene-expression profiles were associated with variability in survival times, 2 separate but complementary approaches were used. In the first approach, the 86 tumors were randomly assigned to equivalent training and testing sets consisting of equal numbers of stage I and III tumors in order to validate a novel risk-index function that captured the effect of many genes at once. In the second approach, cross-validation⁴⁴ was used to more robustly identify the genes associated with survival. Briefly, a 'leave-one-out' cross-validation procedure in which 85 of the 86 tumors (the training set) was used to identify genes that were univariately associated with survival. The risk index was defined as a linear combination of the gene-expression values for the top genes identified by univariate Cox proportional-hazard regression modeling⁴⁵, weighted by their estimated regression coefficients. Kaplan-Meier survival plots and log-rank tests were then used to assess whether the risk-index assignment to high/low categories was validated in the test set. A more detailed description is provided (Supplementary Methods online).

Note: Supplementary information is available on the Nature Medicine website.

Acknowledgments

We thank D. Sanders for technical assistance; D. Sing for assistance with the figures; and G. Omenn for critical reading of this manuscript. This work was supported by National Cancer Institute grant: U19 CA-85953 and the Tissue Core of the University of Michigan Comprehensive Cancer Center (NIH CA-46952).

Competing interests statement

The authors declare that they have no competing financial interests.

RECEIVED 5 APRIL; ACCEPTED 14 JUNE 2002

1. Fry, W.A., Phillips, J.L. & Menck, H.R. Ten-year survey of lung cancer treatments and survival in hospitals in the United States. *Cancer* **86**, 1867–1876 (1999).
2. Williams, D.E. et al. Survival of patients surgically treated for stage I lung cancer. *J. Thorac. Cardiovasc. Surg.* **82**, 70–76 (1981).
3. Pairolero, P.C. et al. Postsurgical stage I bronchogenic carcinoma: Morbid implications of recurrent disease. *Ann. Thorac. Surg.* **38**, 331–338 (1984).
4. Nantke, T. et al. Prognosis and survival in resected carcinoma based on the new international staging system. *J. Thorac. Cardiovasc. Surg.* **96**, 440–447 (1988).
5. Kaisermann, M.C. et al. Evolving features of lung adenocarcinoma in Rio de Janeiro, Brazil. *Oncol. Rep.* **8**, 189–192 (2001).
6. Roggli, V.L. et al. Lung cancer heterogeneity: A blinded and randomized study of 100 consecutive cases. *Hum. Pathol.* **16**, 569–579 (1985).
7. Gail, M.H. et al. Prognostic factors in patients with resected stage I non-small cell lung cancer: A report from the Lung Cancer Study Group. *Cancer* **54**, 1802–1813 (1984).
8. Takise, A. et al. Histopathologic prognostic factors in adenocarcinomas of the peripheral lung less than 2 cm in diameter. *Cancer* **61**, 2083–2088 (1988).
9. Ichinose, Y. et al. Is T factor of the TMN staging system a predominant prognostic factor in pathologic stage I non-small cell lung cancer. *J. Thorac. Cardiovasc. Surg.* **106**, 90–94 (1993).
10. Harpole, D.H. et al. A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression. *Cancer Res.* **55**, 51–56 (1995).
11. Rodenhuis, S. et al. Mutational activation of the K-ras oncogene: A possible pathogenic factor in adenocarcinoma of the lung. *N. Engl. J. Med.* **317**, 929–935 (1987).
12. Slebos, R.J.C. et al. K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung. *N. Engl. J. Med.* **323**, 561–565 (1990).
13. Horio, Y. et al. Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer. *Cancer Res.* **53**, 1–4 (1993).
14. Kern, J.A. et al. C-erbB-2 expression and codon 12 K-ras mutations both predict shortened survival for patients with pulmonary adenocarcinomas. *J. Clin. Invest.* **93**, 516–520 (1994).
15. Ebina, M. et al. Relationship of p53 overexpression and up-regulation of proliferating cell nuclear antigen with the clinical course of non-small cell lung cancer. *Cancer Res.* **54**, 2496–2503 (1994).
16. Mehdi, S.A. et al. Prognostic markers in resected stage I and II non-small cell lung cancer: an analysis of 260 patients with 5 year follow-up. *Clin. Lung Cancer* **1**, 59–67 (1997).
17. Schneider, P.M. et al. Multiple molecular marker testing (p53, c-Ki-ras, c-erbB-2) improves estimation of prognosis in potentially curative resected non-small cell lung cancer. *Br. J. Cancer* **83**, 473–479 (2000).
18. Herbst, R.S. et al. Differential expression of E-cadherin and type IV collagenase genes predicts outcome in patients with stage I non-small cell lung carcinoma. *Clin. Can. Res.* **6**, 790–797 (2000).
19. Liotta, L. & Petricion, E. Molecular profiling of human cancer. *Nature Rev. Genet.* **1**, 48–56 (2000).
20. Golub, T.R. Editorial: Genome-wide views of cancer. *N. Engl. J. Med.* **344**, 601–602 (2001).
21. Bhattacharjee, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
22. Garber, M.E. et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* **98**, 13784–13789 (2001).
23. Mills, N.E. et al. Increased prevalence of K-ras oncogene mutations in lung adenocarcinoma. *Cancer Res.* **55**, 1444–1447 (1995).
24. Giordano T.J. et al. Organ-specific molecular classification of lung, colon and ovarian adenocarcinomas using gene expression profiles. *Am. J. Pathol.* **159**, 1231–1238 (2001).
25. Albertson, D.G. et al. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genet.* **25**, 144–146 (2000).
26. Kansra, S. et al. IGFBP-3 mediates TGF β 1 proliferative response in colon cancer cells. *Int. J. Cancer* **87**, 373–378 (2000).
27. Vadgama J.V. et al. Plasma insulin-like growth factor-I and serum IGF-binding protein 3 can be associated with the progression of breast cancer, and predict the risk of recurrence and the probability of survival in African-American and Hispanic

- women. *Oncology* 57, 330–340 (1999).
28. Voim, M., Mattern, J. & Stammers, G. Up-regulation of heat shock protein 70 in adenocarcinoma of the lung in smokers. *Anticancer Res.* 15, 2607–2609 (1995).
 29. Ciocka, D.R. *et al.* Heat shock protein hsp70 in patients with axillary lymph node-positive breast cancer: prognostic implications. *J. Natl. Cancer. Inst.* 85, 570–574 (1993).
 30. Rotenberg, Z. *et al.* Total lactate dehydrogenase and its isoenzymes in serum of patients with non-small cell lung cancer. *Clin. Chem.* 34, 668–670 (1988).
 31. Krepela, E. *et al.* Cysteine proteases and cysteine protease inhibitors in non-small cell lung cancer. *Neoplasia* 45, 318–331 (1998).
 32. Kos, J. *et al.* Cysteine proteinases and their inhibitors in extracellular fluids: Markers for diagnosis and prognosis in cancer. *Int. J. Biol. Markers* 15, 84–89 (2000).
 33. Golub, T.R. *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537 (1999).
 34. Hedenfalk, I. *et al.* Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539–548 (2001).
 35. Ohta, Y. *et al.* Vascular endothelial growth factor and lymph node metastasis in primary lung cancer. *Br. J. Cancer.* 76, 1041–1045 (1997).
 36. Shibusa, T., Shijubo, N. & Abe, S. Tumor angiogenesis and vascular endothelial growth factor expression in stage I lung adenocarcinoma. *Clin. Cancer Res.* 4, 1483–1487 (1998).
 37. Girardin, S.E. & Yaniv, M. A direct interaction between JNK1 and Crkl is critical for Rac1-induced JNK activation. *EMBO J.* 20, 3437–3446 (2001).
 38. Liu, E. *et al.* The Ras-mitogen-activated protein kinase pathway is critical for the activation of matrix metalloproteinase secretion and the invasiveness in v-crk-transformed 3Y1. *Cancer Res.* 60, 2361–64 (2000).
 39. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* 100, 57–70 (2000).
 40. Hanson, L.A. *et al.* Expression of the glucocorticoid receptor and K-ras genes in urethan-induced mouse lung tumors and transformed cell lines. *Exp. Lung. Res.* 17, 371–387 (1991).
 41. Lin, L. *et al.* A minimal critical region of the 8p22-23 amplicon in esophageal adenocarcinomas defined using STS-amplification mapping and quantitative PCR includes the GATA-4 gene. *Cancer Res.* 60, 1341–1347 (2000).
 42. Kononen, J. *et al.* Tissue microarrays for high throughput molecular profiling of tumor specimens. *Nature Med.* 4, 844–847 (1998).
 43. Johnson, R. & Wichern, D.W. *Applied Multivariate Statistical Analysis*. 543–578 (Prentice Hall, New Jersey, 1988).
 44. Stone, M. Asymptotics for and against cross-validation. *Biometrika* 64, 29–38 (1977).
 45. Cox, D.R. Regression models and life tables. *J.R. Stat. Soc.* 34, 187–220 (1972).

[CANCER RESEARCH 64, 8911-8918, November 15, 2004]

A Novel RAR β Isoform Directed by a Distinct Promoter P3 and Mediated by Retinoic Acid in Breast Cancer Cells

Xinjian Peng,¹ Takeshi Maruo,³ Yanxia Cao,² Vasu Punj,¹ Rajeshwari Mehta,¹ Tapas K. Das Gupta,¹ and Konstantin Christov¹

¹Department of Surgical Oncology, University of Illinois at Chicago, and ²Department of Internal Medicine, Rush University Medical Center, Chicago, Illinois; and ³Department of Obstetrics and Gynecology, Kobe University School of Medicine, Kobe, Japan

ABSTRACT

Retinoids regulate gene transcription through activating retinoic acid receptors (RARs)/retinoic X receptors (RXRs). Of the three RAR receptors (α , β , and γ), RAR β has been considered a tumor suppressor gene. Here, we identified a novel RAR β isoform-RAR β 5 in breast epithelial cells, which could play a negative role in RAR β signaling. Similar to RAR β 2, the first exon (59 bp) of RAR β 5 is RAR β 5 isoform specific, whereas the other exons are common to all of the RAR β isoforms. The first exon of RAR β 5 does not contain any translation start codon, and therefore its protein translation begins at an internal methionine codon of RAR β 2, lacking the A, B, and part of C domain of RAR β 2. RAR β 5 protein was preferentially expressed in estrogen receptor-negative breast cancer cells and normal breast epithelial cells that are relatively resistant to retinoids, whereas estrogen receptor-positive cells that did not express detectable RAR β 5 protein were sensitive to retinoid treatment, suggesting that this isoform may affect the cellular response to retinoids. RAR β 5 isoform is unique among all of the RARs, because a corresponding isoform was not detectable for either RAR α or RAR γ . RAR β 5 mRNA was variably expressed in normal and cancerous breast epithelial cells. Its transcription was under the control of a distinct promoter P3, which can be activated by all-*trans*-retinoic acid (atRA) and other RAR/RXR selective retinoids in MCF-7 and T47D breast cancer cells. We mapped the RAR β 5 promoter and found a region -382/-99 to be the target region of atRA. In conclusion, we identified and initially characterized RAR β 5 in normal, premalignant, and malignant breast epithelial cells. RAR β 5 may serve as a potential target of retinoids in prevention and therapy studies.

INTRODUCTION

The biological effects of retinoids are mainly mediated by two families of nuclear receptors: retinoic acid receptors (RARs) and retinoic X receptors (RXRs), each consisting of three receptor subtypes (α , β , γ ; refs. 1, 2). In addition, each RAR gene generates multiple isoforms by either alternative splicing or differential usage of two promoters (1, 2). RARs/RXRs belong to the superfamily of two promoters (1, 2). RARs/RXRs belong to the superfamily of nuclear receptors that mediate the transcriptional effects of steroid hormones, vitamin D, and thyroid hormone (3). RARs preferentially dimerize with RXRs to form RAR-RXR heterodimers that are thought to be obligatory intermediates in the effects of RAR ligands on gene expression (4). RXRs also can homodimerize to form transcriptionally active complexes (5). Homo- and heterodimeric retinoid receptor complexes bind to distinct retinoid response elements embedded in the regulatory regions of retinoid-responsive genes (6). Although there is considerable variability in the sequence and structure of the

retinoid response elements in retinoid-regulated genes, they conform to a general canonical sequence in which two directly repeated receptor-binding hexanucleotide motifs [consensus (A/G)G(G/T)TCA] are separated by a variable number of intervening nucleotides (6).

RAR β itself is a retinoid target gene and believed to play a role as a tumor suppressor gene in tumorigenesis (7). The human RAR β gene was first identified from hepatocellular carcinoma in 1987 (8), followed by the identification of retinoic acid response element (RARE) in its promoter region (9). In the mice, the RAR β gene generates four distinct transcripts: splice variants RAR β 1 and RAR β 3 from transcription at promoter P1, and RAR β 2 and RAR β 4 from the RAR β -containing P2 promoter (2, 10). In the human, only RAR β 2 and RAR β 4 transcripts have been identified in normal adult cells (11). Human RAR β 1 is expressed in fetal tissues and some small cell lung carcinoma cell lines (12); whereas a human homologue of the RAR β 3 isoform has not been detected (7). The RAR β 2 and RAR β 4 transcripts differ only in the content of their 5'-most exon, a result of alternative splicing (10). On the basis of homology with other members of the steroid hormone receptor superfamily, six distinct domains (A-F) have been identified within RARs and RXRs (2). Thus far, all of the identified RAR isoforms are only different at their unique A domain and are derived from two promoters and alternative splicing (11). Isoforms of a given RAR gene generally contain identical protein sequences B-F (11).

We've been interested in RAR/RXR alteration in the process of breast tumor progression with a MCF10 model (13). During the characterization of RAR β expression in the MCF10 series of cell lines (benign MCF10A, premalignant MCF10AT, and malignant MCF10CA1a cell lines; ref. 14), we identified a novel RAR β isoform, which we named RAR β 5. RAR β 5 mRNA expression is under the control of a distinct promoter P3 and is mediated by all-*trans*-retinoic acid (atRA) and other RAR/RXR selective ligands in breast cancer cells. It was detected in both normal and breast cancer cells. We also cloned and initially characterized the promoter region of RAR β 5. The same protein isoform was previously associated with RAR β 4 transcript (11, 15) and then termed RAR β ' (7). In this study, we have identified RAR β 5 at both gene and protein level.

MATERIALS AND METHODS

Cell Culture. The MCF10A cell line was received from Karmanos Cancer Institute (Detroit, MI) and cultured as described previously (13). Normal human mammary epithelial cells (HMEC) were purchased from Clonetics (Santa Rosa, CA) and cultured in MEGM with supplements (Clonetics). MCF-7, T47D, and MDA-MB435 cell lines were purchased from American Type Tissue Collection (Manassas, VA). BCA-1 to BCA-11 breast carcinoma cell lines were from breast cancer patients and established in our laboratory (16). These cell lines are still at their early passages (passage number at 7-10), and their characteristic features are summarized in Appendix 1 as supplemental material. All of these cell lines were cultured in MEM supplemented with 100 units/mL penicillin, 100 μ g/mL streptomycin, and 10% fetal bovine serum, 200 μ mol/L L-glutamine and 100 μ mol/L MEM nonessential amino acids. The atRA was purchased from Sigma (St. Louis, MO). The 9-*cis*-retinoic acid (9-*cis*RA), 4-hydroxyphenyl retinamide (4-HPR), and LGD1069 were obtained from the repository of the National Cancer Institute (Bethesda, MD).

Received 5/23/04; revised 9/14/04; accepted 10/14/04.

Grant support: The United States Army Breast Cancer Research Program DAMD17-99-9221 (K. Christov) and the Illinois Department of Public Health Penny Savana Breast and Cervical Cancer Research Fund (X. Peng).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article may be found at Cancer Research Online at <http://cancerres.aacrjournals.org>. The sequences reported in this article have been deposited in the GenBank database (accession nos. AY501390-AY501391).

Requests for reprints: Konstantin Christov, Department of Surgical Oncology, University of Illinois at Chicago, 840 South Wood Street (MC 820), Chicago, IL 60612. Phone: (312) 996-5347; Fax: (312) 996-9365; E-mail: christov@uic.edu.

©2004 American Association for Cancer Research.

A NOVEL RAR β ISOFORM IN BREAST CANCER CELLS

Am80 (RAR α / β selective ligand) was a generous gift from Dr. Koichi Shudo (ITSUU Laboratory, Tokyo, Japan).

Rapid Amplification of cDNA 5'-Ends and cDNA Cloning. Rapid amplification of cDNA 5'-ends (5'-RACE) was done with SMART RACE cDNA Amplification Kit (Clontech Laboratories, Inc., Palo Alto, CA) according to the User Manual. RACE PCR was done with a Universal Primer and two RAR β gene-specific reverse primers located at 1754 bp (RAR β -1754RP, GGACTGTGCTCTGCTGTGTTCCCACTT) and 1216 bp (RAR β -1216RP, GGTCTGCGATGGTCAAGCCAGTGAA) 3' of the RAR β transcription site. PCR products were cloned into pCR4-TOPO vector (Invitrogen, Carlsbad, CA) and sequenced with ABI PRISM 377 DNA Sequencer (Applied Biosystems, Foster City, CA).

Reverse Transcription-PCR. RT was done in a final volume of 20 μ L with 2 μ g total RNA and 100 units of MuLV reverse transcriptase (Invitrogen) at 42°C for 50 minutes. Conventional PCR was mainly used to qualitatively detect gene expression. PCR was done with 1 μ L RT product with PCR Supermix (Invitrogen). PCR primer pairs are RAR β (475FP, GACTGTAT-GGATGTTCTGTGTCAG; and 730RP, ATTTGTCTGCGACAGCAAGCA) and RAR β (14FP, CTGGAAGTCTGTACACAGTGA; and 343RP, GGA-CATTCCTCACTTC-AAAGC). β -actin (FP, GTCAACCAACTGGGACGACA; and RP, TGGCCATCTCTTGC-TCGAA) was used as an internal control. Real-time PCR was done with 1 μ L RT product with 7900HT Sequence Detection System (ABI, Applied Biosystems) and ABI 2 \times SYBR Green PCR Master Mix (ABI# 4309155) according to recommended guidelines of ABI. Primer pairs for real-time PCR were RAR β (584FP, GATTGACCCAAAC-GGAATGGCAGCA; and 730RP) and RAR β (15FP, GGAAGGT-CGTACA-CAGTGAATTCTCTGAG; and RAR β -730RP); real-time PCR data were analyzed with a software package (ABI Prism SDS2.1) provided with the instrumentation system.

Expression Vector and *In vitro* Translation. The RAR β expression vector was generated by PCR-cloning with pcDNA3.1/V5-His TOPO TA Expression Kit (Invitrogen). The open reading frame (ORF) of RAR β was isolated by PCR of 5'-RACE cDNA from MDA-MB435 cells with primers containing RAR β start and stop codon (RAR β -start, CAGAAGAATAT-GATTTACACTTGTACCCG; and RAR β -stop, GTCTTATTCACCGA-GT-GGTGACTG). RAR β expression vector pTag (RAR β 5'-; RAR β with a mutation to knockout a downstream translation start site) was a generous gift from Dr. Karen Swisschelm (Department of Pathology, University of Washington, Seattle, WA; ref. 11). RAR β insert was cut from pTag (RAR β 5'-) and ligated into the *Bam*HI sites of pcDNA3.1 vector (Invitrogen) to generate RAR β expression vector pcDNA3.1(RAR β). The pcDNA3.1(RAR β) was used for both *in vitro* translation to generate RAR β protein as positive control for Western blot and cotransfection to test the effects of RAR β expression on RAR β promoter activity. *In vitro* translation was done with TNT Quick Coupled Translation kit (Promega, Madison, WI).

Western Blot. When cells grew to 50 to 70% confluence, cell lysates were prepared and subjected to Western blot analysis as described previously (17). Two antibodies were used to detect RAR β isoforms, one recognizing amino acids 430-447 in the COOH-terminus of RAR β (sc-552, Santa Cruz Biotechnology Inc., Santa Cruz, CA), the other one recognizing amino acids 407-423 in the COOH-terminus of RAR β (Geneka Biotechnology Inc., Montreal, Quebec). Another two antibodies recognizing COOH-terminus of RAR α (sc-551, Santa Cruz Biotechnology Inc.) and RAR γ (sc-550, Santa Cruz Biotechnology Inc.) were used to detect RAR α and RAR γ isoforms.

The 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium Bromide Assay for Cell Growth. Cell proliferation was examined by colorimetric [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; MTT] assay. MTT is a pale yellow substrate that is cleaved by living cells to yield a dark blue formazan product. This process requires active mitochondria, and even freshly dead cells do not cleave substantial amounts of MTT. Briefly, cells (500 cells per well) were seeded in 96-well plates and cultured overnight. Then cells were incubated with 1 μ mol/L retinoids, and the media were changed every second day. After 7-day treatment, 0.01 mL of MTT solution (5 mg/mL) was added to each well, mixed gently, and incubated with the cells at 37°C for 2 to 3 hours. The media were carefully removed, 0.1 mL of DMSO was added to each well, and plates were assayed for cell proliferation as described previously (18).

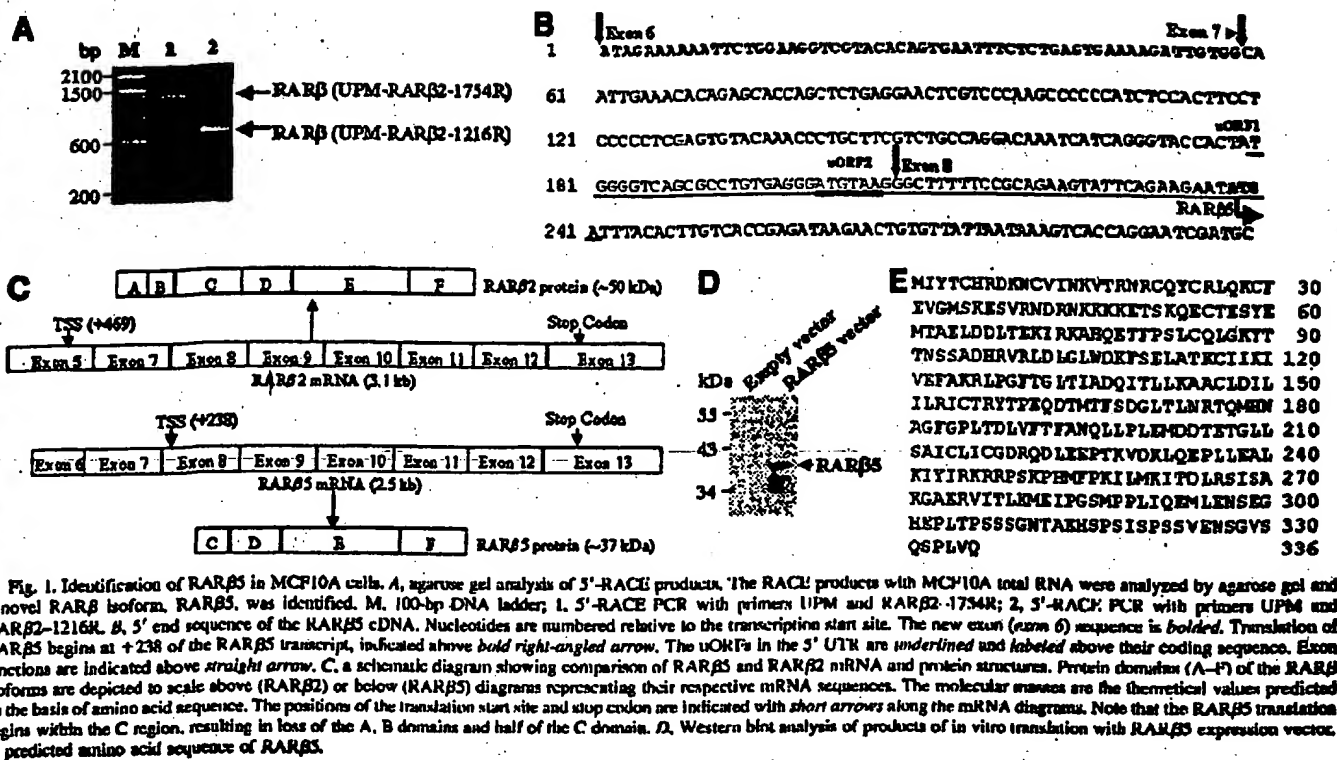
RAR β Promoter-Luciferase Reporter Plasmids. The 1-kb 5' flanking region (P-1000/+33 relative to the transcription start site) of RAR β was first

isolated by PCR from genomic DNA extracted from MDA-MB435 breast cancer cells with Advantage 2 PCR kit (Clontech). Primer pairs were designed to contain a *Kpn*I restriction site at the 5' end of the forward primer and a *Xho*I restriction site at the 5' end of the reverse primer. The PCR product was first cloned to the pCR4-TOPO vector, and then subcloned to the *Kpn*I/*Xho*I sites of the promoterless PGL3 basic vector (Promega). Orientation and sequence of all of the constructs were verified by direct sequencing. All of the other promoter deletion mutation constructs (P-428/+33, -323/+33, -302/+33, -177/+33, -99/+33) were cloned in the same way with PGL3-P-1000/+33 as a template.

Cell Transfections and Luciferase Assay. MCF-7 and T47D cells were plated at 1 to 1.2 \times 10⁵ cells per well in 12-well plates. After overnight incubation, the media were replaced by MEM containing 2% fetal bovine serum. Transient transfection was done in the same media with Lipofectamine 2000 (Invitrogen). Cells were transfected with 0.5 μ g/well promoter constructs (0.5 μ g/well for PGL3-P-1000/+33, the amount of other deletion mutants was correspondingly adjusted to make each well contain the same amount of the plasmids) with or without 0.5 μ g/well pcDNA3.1 empty vector or 0.64 μ g/well pcDNA3.1(RAR β) expression vector. A 20 ng/well pCMV β gal vector (Clontech) was cotransfected as an internal control for transfection efficiency. After 5 hours incubation, the medium was replaced with a fresh one containing 1 μ mol/L *at*RA or other retinoids or DMSO (solvent control, 1 μ L/10 mL media), and cells were incubated for an additional 24 hours. Luciferase and β -galactosidase activities were assayed with Luciferase Reporter Assay Kit and Luminescent β -gal detection kit II (Clontech).

RESULTS

A Novel RAR β Transcript in MCF10A Breast Epithelial Cells. During characterization of RAR β expression in MCF10A series of cell lines (14), with primers recognizing RAR β coding region, we detected RAR β transcript by reverse transcription (RT)-PCR; but we failed to detect RAR β transcript with primers recognizing RAR β 5'-untranslated region (UTR). Hence, to examine the 5' region of RAR β in MCF10A cells, we did a 5'-RACE analysis of the MCF10A total RNA with two RAR β specific primers (Fig. 1A). Using these two primers, we failed to detect the expected RAR β fragments with size ~1.8 kb and ~1.3 kb, respectively, but did consistently detect a band with a smaller size (~0.6 kb shorter). The 5'-RACE product was cloned and sequenced. The Blast search of GenBank suggests it to be a novel RAR β isoform that has not yet been reported, henceforth it is referred to as RAR β 5. The 5' end sequence of RAR β 5 cDNA is presented in Fig. 1B. Only the first exon (Fig. 1; exon 6, 58 bp, *hnd*) is RAR β 5 specific. We also used another primer set designed in terms of the RAR β 5-specific sequence (RAR β 5-14FP) and the 3'-UTR sequence of RAR β (RAR β -1827RP) to amplify a fragment spanning the whole coding region of RAR β 5. Cloning and sequencing analysis of this PCR product showed a unique first exon of the RAR β 5, whereas all of the downstream exons are common to all of the isoforms of RAR β . Alignment of the first exon of RAR β 5 to bacterial artificial chromosome (BAC) clones RP11-421F9 and RP11-733H11 (GenBank accession nos. AC133141.2 and AC098477.2) shows that the first exon of RAR β 5 is located ~29.4 kb downstream of the first exon (exon 5) of RAR β 2 and ~2.8 kb upstream of the second exon of RAR β 2. Therefore, this novel exon is numbered exon 6, and numeration for the downstream exons is updated from what was reported previously (ref. 11; Fig. 1C). The 5'-UTR of RAR β 5 mRNA is 237 nucleotides long and contains 2 upstream ORFs (uORFs; Fig. 1B). We note in this respect that both RAR β 2 and RAR β 4 contain multiple uORFs (8, 11) that could play a role in tightly controlling translation efficiency. Differing from RAR β 2, the first exon of RAR β 5 does not contain any translation start codon, the 5'-most AUG of the RAR β 5 transcript with the RAR β 5 coding sequence is located at nucleotide 238 of RAR β 5 mRNA (within the third exon of RAR β 5) and corresponds to an internal methionine codon at amino acid 113 of the RAR β 2 protein

A NOVEL RAR β ISOFORM IN BREAST CANCER CELLS

(Fig. 1, B and C). This AUG is within an appropriate nucleotide context for translation initiation (19) and would result in a protein of 336 amino acids with an estimated molecular mass of ~37 kDa (Fig. 1C). *In vitro* translation of RAR β 5 expression vector confirmed that this AUG is a functional translation initiation codon (Fig. 1D). It should be noted that *in vitro* translation generated multiple protein bands, which is consistent with a previous report and might not happen in the cells (11), probably because of lack of the whole UTR region in the expression vector and lack of natural chromatin environment. This RAR β 5 protein product is identical to a truncated RAR β 2 or RAR β 4 protein reported previously (7, 11). The predicted amino acid sequence is given in Fig. 1E.

AtRA Mediated Expression and Regulation of RAR β 5. Identification of RAR β 5 raised a question as to whether its expression is mediated by atRA. To directly examine the presence of RAR β 5 in comparison with RAR β 2 in patients, we used the human breast cancer cells derived from the patients and being in early *in vitro* (<10) passages. In addition, we also examined the expression of both RAR β isoforms in established breast cancer cell lines and normal human mammary epithelial cells. RT-PCR analysis showed that all of the examined breast cancer cells expressed detectable RAR β 5 mRNA, but its expression was differentially regulated by atRA treatment (Fig. 2A). It was up-regulated by atRA in BCA-1, 3, and 4 cells, whereas BCA-8 showed a slight down-regulation of RAR β 5 by atRA. Similarly, the level of RAR β 2 was up-regulated by atRA in some cells (BCA-1, 3, 4, 8, 9, and 10), whereas in others it remained unaltered. However, no correlation between RAR β 2 and RAR β 5 expression could be established. Since in breast cancer, estrogen receptor (ER) plays a critical role in its response to various chemotherapeutic agents and to quantitate the expression of RAR β 5 and RAR β 2 mRNA, we did real-time PCR with HMEC, ER-positive breast cancer cell lines MCF-7 and T47D, and the ER-negative breast cancer cell line MDA-MB435 that expresses a high level of RAR β 2 mRNA (11). Real-time PCR clearly showed that RAR β 5 was preferentially up-regulated by atRA in MCF-7 cells, whereas RAR β 2 was preferentially up-

regulated by atRA in T47D cells (Fig. 2B). ER-negative MDA-MB435 cells expressed a high level of RAR β 2 but a low level of RAR β 5 relative to HMEC cells. RAR β 5 and β 2 were consistently expressed at a low level in ER-positive MCF-7 and T47D cells relative to normal HMEC cells (Fig. 2B). Because other retinoids function through a similar mechanism in the cells, it is reasonable to expect that other retinoids might also mediate RAR β 5 expression. By

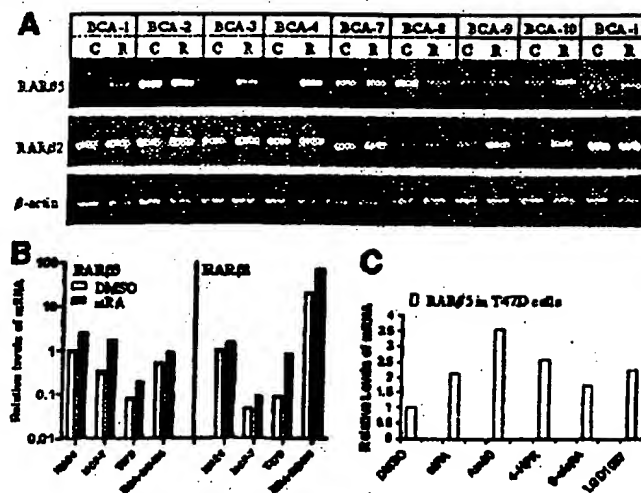


Fig. 2. RT-PCR analysis of RAR β 5 and RAR β 2 mRNA expression and regulation by retinoids. A, Cells were treated with 1 μ M atRA for two days, and total RNA was subjected to RT-PCR analysis with specific primers. Both RAR β 5 and RAR β 2 are differentially expressed in all of these tumor cells and regulated by atRA. B, Real-time PCR analysis showing relative levels of RAR β 5 and RAR β 2 mRNA normalized to β -actin (the basal RAR β mRNA level in HMEC is set as 1) after atRA treatment (1 μ M atRA for 24 hours) in HMEC and breast cancer cell lines. Results are expressed as the mean value of two independent experiments. C, Real-time PCR analysis showing relative levels of RAR β 5 mRNA normalized to β -actin after treatment with RAR/RXR selective ligands (1 μ M for 24 hours) in T47D cells. The atRA served as a positive control. Results are expressed as the mean value of duplicate analyses of the same cDNA samples.

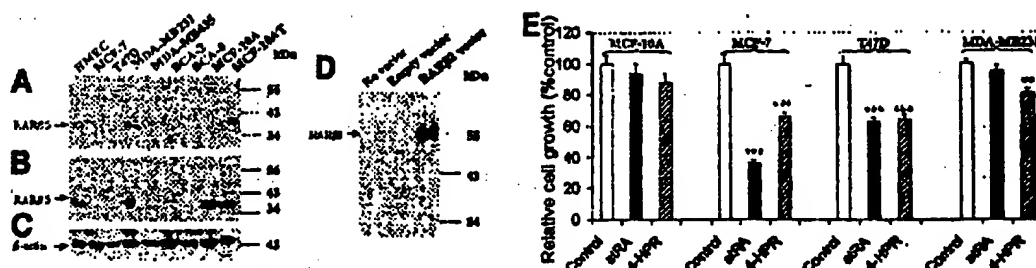
A NOVEL RAR β ISOFORM IN BREAST CANCER CELLS

Fig. 3. Detection of endogenous levels of RAR β 5 protein in normal and breast cancer cells and cellular sensitivity to retinoid treatment. A and B. Western blot analysis of cell extracts of various cells with RAR β specific antibodies. Twenty micrograms (A) and 60 μ g (B) of the total proteins from each cell line were loaded for immunoblot analysis with two polyclonal antibodies raised against different COOH-terminal RAR β epitopes [A, antibody recognizing amino acids 430-447 of RAR β (sc-352, Santa Cruz Biotechnology); and B, antibody recognizing amino acids 407-423 of RAR β (16021053, Genetex)], respectively. RAR β 5 protein was detected as a ~37 kDa protein band. C. β -actin was used as an internal control. D. Western blot analysis of products of *in vitro* translation with different vectors. The positions of molecular mass markers are indicated to the right. RAR β 2 (~55 kDa) was not detectable in any of the cell lines except positive control (D). E. MTT assay of cell proliferation in response to retinoids. Data are expressed as the percentage of DMSO control \pm SD of 8 wells. All of the data shown are representative of three independent experiments. **, $P < 0.01$ compared with control; ***, $P < 0.001$ compared with control.

using conventional RT-PCR, we confirmed that 4-HPR and 9-*cis*RA also differentially up-regulated RAR β 5 expression in MCF-7, T47D, and BCA-3 cells. Fig. 2C presents data showing real-time RT-PCR analysis of RAR β 5 expression mediated by different RAR/RXR selective ligands in T47D cells. In addition to *at*RA, Am80 (RAR β / α selective ligand) and LGD1069 (RXR selective ligand) also significantly up-regulated RAR β 5 expression, whereas 4-HPR (a weak RAR γ ligand) and 9-*cis*RA (RAR/RXR ligand) had relatively low efficacy in mediating RAR β 5 expression (Fig. 2C).

RAR β 5 Protein Expression in Correlation to Cellular Resistance to *at*RA. We did Western blot to analyze RAR β protein expression in a panel of breast epithelial cells (normal HMEC; ER-positive MCF7 and T47D; ER-negative MDA-MB231, MDA-MB435, human breast carcinoma BCA-2 and BCA-8, MCF10A benign, and MCF10AT premalignant breast epithelial cells) with two RAR β polyclonal antibodies that were raised against amino acids at the COOH-terminal (a region common among human RAR β isoforms). A protein band with the expected molecular mass (~37 kDa) was detected in HMEC, MDA-MB231, BCA-2, MCF10A, and MCF10AT cells by both antibodies (Fig. 3, A and B), suggesting this protein could be RAR β 5. An RAR β 2 protein band (~55 kDa) was only detected from the *in vitro* translation product (Fig. 3D). None of the cell lines expressed detectable RAR β 2. In another experiment, we were not able to detect RAR β 2 protein expression in all of the BCA (-1, -2, ..., -11) cell lines, but we detected RAR β 2 protein in HMEC cells that were from a different source (Cambrex Bio Science Inc., Walkersville, MA)⁴ with the same antibody (C-19, Santa Cruz Biotechnology), indicating that RAR β 2 protein expression could be cell-type specific. The ~37 kDa protein detected in this study could be identical to the RAR β protein isoform (termed RAR β 4, ~40 kDa) identified previously in breast cancer cells (11), because same antibody was used for the detection, and the molecular size is also very close considering the 10% margin of error for our molecular mass standards. This RAR β protein isoform seemed to be preferentially expressed in ER-negative normal and cancerous breast epithelial cells, but it was not detectable in ER-positive breast cancer line MCF-7 and T47D.

To evaluate whether RAR β 5 expression is associated with cellular resistance to retinoids, cell lines expressing different level of RAR β 5 protein (Fig. 3, A and B) were selected to assess their sensitivity to retinoids with MTT assay. Immortalized benign MCF10A cells, which express high level of RAR β 5, were resistant to both *at*RA and 4-HPR; ER-negative MDA-MB231 cells, which also express RAR β 5 protein, showed resistance to *at*RA, but 4-HPR effectively inhibited

the proliferation of MDA-MB231 cells. ER-positive MCF-7 and T47D cells in which RAR β 5 protein expression was not detectable were sensitive to both *at*RA and 4-HPR (Fig. 3E). In addition, MDA-MB435, MCF10AT, and BCA-2 cells, which express detectable RAR β 5 protein, were relatively resistant to *at*RA (data now shown). These results suggest that RAR β 5 might contribute to cellular resistance to *at*RA, which functions through receptor-dependent pathway. RAR β 5 did not have much influence on the effect of 4-HPR, which functions through both receptor-dependent and independent pathway (20).

Sequence Analysis of 5' Flanking Region of RAR β 5. Although the RAR β 5 regulation pattern by *at*RA was more or less similar to that of RAR β 2, in some cells such as BCA-2 and BCA-8, MCF-7 and T47D, the expression pattern was significantly different. Sequence alignment showed that the first exon of RAR β 5 is far away (~30 kb) from the P2 promoter, suggesting that RAR β 5 and RAR β 2 use different promoters. Therefore, we cloned and sequenced the 1-kb 5' flanking region of RAR β 5. Fig. 4 shows the 480-bp 5' flanking sequence of RAR β 5. The 5' flanking sequence (-1000/-59) of RAR β 5 was analyzed with several promoter identification programs including Proscan (21), Promoter 2.0 (22), BDGP: Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html), McPromoter (<http://genes.mit.edu/McPromoter.html>), PromoterInspector (<http://www.genomatix.de/>). None of these programs was able to predict this promoter. The region close to the putative transcription start site lacks the canonical TATA and CCAAT boxes, but a TATA-like box (TATAATT) is present 42 bp upstream of the transcription start site. Additional analysis with MatInspector (23) and TFSEARCH

```

-480 AGTACGACCTTAAACACACACCTCCACCTTCTCTCTGTGAGAACTCAATGCGAT
      CDP      AP1
-420 TTTATAGTGTGATATGACTCTGTCTGTTTGTGAGACATGATGAGTAA
      NFY1      RXR
-360 GCACACAATCAACCACTCCAAAGGTTTGAGAGGATCTTAACTTTAG
      GATA      GATA
-300 TTTTCAGGAGTGTAAATCTACACACCTTTCTAGAGTCACTTTAACTGAT
      SRY      CEBP/β
-240 TAAATTTAATATTTTAAATTAACAACTTTGAACTTTTACACITCTGCT
      NFAT      CCT1
-180 ATTGTGACCTTAAGATGTGCTGAGATGACTACAAATTTGTCATCAATGGA
      GATA      NFAT      AP1
-120 TTTCTGTGAGAAATATTTTGAAGTAAATTTGATATCAATGATTA
      AP1      SRY
-60 AGTCTCTTTGTCTGATATATTTTCACTTTAGTAGACAAACCAACATTTATGTC
+1 ATGAGAAAAATTTCTGAGGTCGTACACAGTGAATTTCTGTGAGTGAAGATTTGTC

```

Fig. 4. Nucleotide sequence of the 5' flanking region of human RAR β 5 gene. Transcription start site is underlined in bold. Shading denotes the core sequence of potential transcription binding sites with high identity to authentic core and matrix sequences as identified by MatInspector V2.2. Nucleotides are numbered negatively to the left of the sequence with nucleotide +1 corresponding to the transcription start site. A TATA-like consensus sequence is boxed.

⁴ X. Peng, D. Yan, K. Christov, unpublished data.

A NOVEL RAR β ISOFORM IN BREAST CANCER CELLSTable 1 Transcriptional initiation site mapping (5'-RACE) of the hRAR β gene in breast epithelial cells

hRAR β transcriptional start site sequence		
Clone	MCF10A	MDA-MB435
1	ATAGAAAAAAT	ATAGAAAAAAT
2	ATAGAAAAAAT	ATAGAAAAAAT
3	ATAGAAAAAAT	ATAGAAAAAAT
4	ATAGAAAAAAT	ATAGAAAAAAT
5	ATAGAAAAAAT	ATAGAAAAAAT
6	ATAGAAAAAAT	ATAGAAAAAAT

(<http://www.cbrc.jp/research/db/TFSEARCH.html>) identified DNA binding sites for AP-1, GATA, SRY, CEBP β , NFAT, OCT1, and so forth at the region -500/+1 (Fig. 4). In addition, Promo (24) identified a binding site for RXR α at 312 bp upstream of the transcription start site. However, these are the potential binding sites, and tests for functionality need to be done to confirm their importance.

Transcription Start Site of RAR β . To determine the transcription start site of RAR β , we did 5'-RACE analysis on mRNA from MDA-MB435 and MCF10A cells. The 5' regions of both RAR β 2 and RAR β 5 were cloned from MDA-MB435 cells and sequenced. Because only a single transcription start site in the human RAR β 2 gene has been defined (9), RAR β 2 cDNA cloning and sequencing served as a good control for mapping the transcription start site of RAR β 5 with 5'-RACE analysis. In total, six RAR β 5-positive clones from MCF10A 5'-RACE products and four RAR β 5-positive and two RAR β 2-positive clones from MDA-MB435 were sequenced. Sequence analysis showed a single transcription start site of RAR β 5 in both MCF10A and MDA-MB435 cells (Table 1). Interestingly, a single nucleotide (A) deletion close to transcription start site was found in five of the six clones from MCF10A cells and in all four clones from MDA-MB435 cells. However, as no deletion was found in the corresponding DNA region in MDA-MB435 cells, it seems that the deletion happened during the transcription process. Two transcription start sites [+1 and -11; +1 identifies the first nucleotide of the putative transcription start site based on GenBank sequence data (NM_000965)] of RAR β 2 were identified in MDA-MB435 cells; it seems that RAR β 2 has multiple transcription start sites in this cancer cell line.

RAR β 5 Promoter Activity. A series of RAR β 5 promoter-luciferase reporter vectors were constructed. When these constructs were transfected into MCF-7 and T47D cells (which are ER-positive and responsive to aTRA) and assayed for reporter gene activity, although differential promoter activity was observed in the two cell lines, the region -99/+33 consistently showed negligible or very low promoter activity in the two cell lines, suggesting either the existence of a negative regulatory element within this region or the presence of a strong activator in the region between -177 and -99 (Fig. 5, A and B). In MCF-7 cells, PGL3-1000/+33, 428/+33, -323/+33, -302/+33, and -177/+33 exhibited significant basal promoter activity (relative to the empty PGL3 Basic vector control). The aTRA treatment additionally increased promoter activity by 2- to 5-fold in MCF-7 cells but only 2- to 3-fold in T47D cells, which was in agreement with real-time PCR results (Fig. 2B). Deletion of region -302/-177 significantly decreased promoter activity induced by aTRA in MCF-7 cells. On the basis of transfection assay data from both cell lines, it seems that the promoter region -302/-99 is the target region for aTRA stimulation, whereas no RARE/RXRE was identified in this region. Therefore, either the target binding site has not been identified, or the stimulatory effect caused by aTRA is an indirect effect.

Because aTRA functions through receptor-dependent pathway, we hypothesized that expression of RAR β 2 could affect RAR β 5 promoter activity mediated by aTRA. To test this hypothesis, RAR β 2 expression vector (pcDNA3.1-RAR β 2) was cotransfected with RAR β 5 promoter constructs into MCF-7 and T47D cells. After transfection, the cells were incubated in the presence or absence of aTRA for 24 hours, and luciferase assay was done for promoter activity. Surprisingly, cotransfection of the empty vector pcDNA3.1 also greatly increased RAR β 5 promoter activity, whereas cotransfection of RAR β 2 expression vector pcDNA3.1-RAR β 2 did not cause a significant increase in promoter activity relative to the control (Fig. 5, C and D); however, in the presence of aTRA, RAR β 2 expression did significantly increase promoter activity compared with that of empty vector-transfected cells in MCF-7 cells (Fig. 5C), suggesting that activation of RAR β 5 promoter activity by aTRA is receptor dependent.

We also examined RAR β 5 promoter activity in the presence of other RAR/RXR selective ligands in both MCF-7 and T47D cells. As

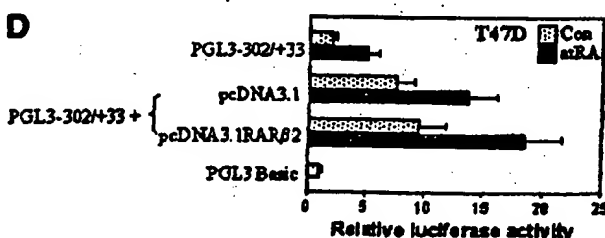
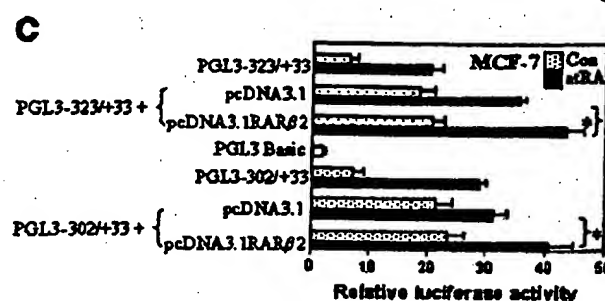
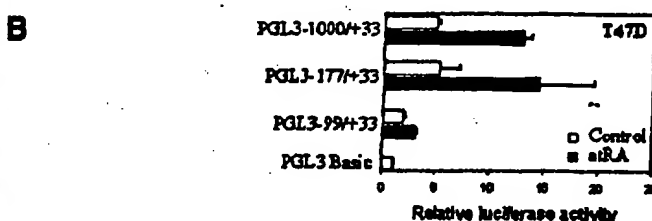
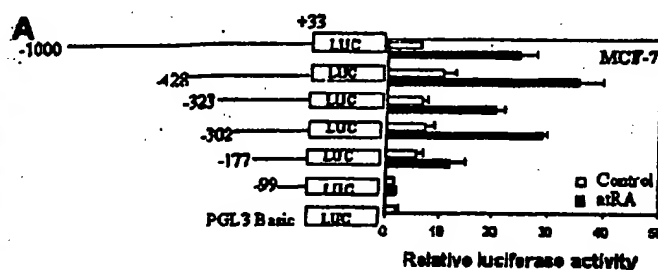


Fig. 5. RAR β 5 promoter activity in MCF-7 and T47D cells. A and B, a region -302/-99 was found to be the target region for aTRA-induced promoter activity. RAR β 5 promoter activity was assayed in MCF-7 (A) and T47D (B) cells transfected with deletion mutants in the 5' region. C and D, effects of cotransfection of RAR β 2 expression vector on promoter activity in MCF-7 (C) and T47D (D) cells. Schematic representations of the 5' deletion constructs are shown to the left of the graph (A). Results are of three independent experiments done in triplicate; bars, mean \pm SEM. Relative luciferase activity, luciferase activity normalized to β -galactosidase. *, $P < 0.05$ compared with the corresponding control.

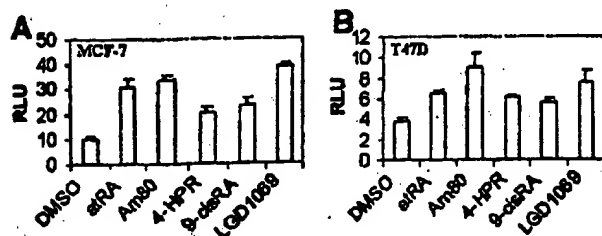
A NOVEL RAR β ISOFORM IN BREAST CANCER CELLS

Fig. 6. RAR β promoter activity is up-regulated by various RAR/RXR selective ligands in MCF-7 and T47D cells. Cells were transfected with PGL3-1000/+33-RAR β promoter construct and treated with 1 μ mol/L retinoids for 24 hours. Results are from triplicate wells of one experiment; bars, mean \pm SEM. RLU, relative luciferase activity normalized to β -gal. The atRA treatment served as a positive control.

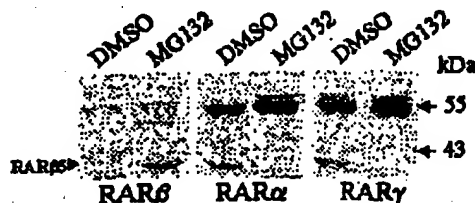


Fig. 7. RAR β is unique among all of the RARs and not a cleaved product from RAR β . Western blot analysis of RAR β , RAR α , and RAR γ in MCF-7 cells. MCF-7 cells were treated with DMSO (control) or MG132 (50 μ mol/L) for 5 hours, cell lysates (50 μ g) were subjected to Western blot analysis with polyclonal antibodies against RAR β (Santa Cruz Biotechnology, sc-552), RAR α (Santa Cruz Biotechnology, sc-551), and RAR γ (Santa Cruz Biotechnology, sc-550).

shown in Fig. 6, A and B, all of the tested RAR/RXR selective ligands differentially increased RAR β promoter activity, whereas Am80 showed the highest efficacy. These data also confirmed our RT-PCR analysis (Fig. 2C).

Is RAR β a Unique Isoform among RARs? The identification of RAR β raises the question as to whether corresponding isoform exists in the other two RARs. Sequence analysis revealed the possibility of existence of a corresponding isoform in both RAR α and RAR γ , because a similar internal start codon is present in both RAR α and RAR γ at similar positions. In addition, we also observed a similar protein band at the position of ~37 kDa in MCF-7 and MDA-MB231 cells through Western blot analysis with RAR α - and RAR γ -specific antibodies that are not cross-reactive with each other and recognize the COOH-terminus of the corresponding RAR isoform. Therefore, we first 5'-RACE-analyzed the mRNA extracted from MCF-7 and MDA-MB231 cells with RAR α -specific primers but failed to detect any expected cDNA band with a smaller size corresponding to the putative truncated RAR α . Only a single RAR α band of an expected size corresponding to full-length RAR α was detected (data not shown). Because these receptor isoforms degrade quickly, we therefore hypothesized that the observed protein band of low molecular size for RAR α and RAR γ might be a fragment generated from protease cleavage. We used cell permeable proteasome inhibitor MG132, which can inhibit the degradation of RAR α and RAR γ (25). MG132 treatment completely blocked the generation of this fragment (Fig. 7), showing that these bands are products of protease cleavage of RAR α and RAR γ . The biochemical and biological properties of these RAR α and RAR γ fragments are not clear at present. When cells were treated with MG132, a corresponding equivalent band of RAR β was observed. Whether the expression of RAR β protein is because of inhibition of protein degradation or induction by MG132 treatment still needs additional in-depth studies.

Because a major level of regulatory control by retinoids is post-translational, we examined effect of atRA treatment on RAR β protein expression with and without proteasomal inhibition. Because proteasomal inhibitors are generally cytotoxic, a period of 8.5 hours

was found not to trigger significant cell death in the cells examined. Cells were treated for 8 and 24 hours with 1 μ mol/L atRA, and then treated with or without 40 μ mol/L MG132 for the final 8.5 hours. We did not observe significant alteration of RAR β protein level in MDA-MB435 cells; however, no RAR β protein was detectable in T47D cells in either condition (data not shown). Nevertheless, MG132 effectively blocked atRA-induced RAR α degradation in both cell lines as observed in MCF-7 cells (25).

Genomic Structure of RAR β in Comparison to RAR β 2. On the basis of the identified RAR β cDNA sequence, the known RAR β 2 sequence, and the published Human Genome Project Data (<http://www.ncbi.nlm.nih.gov/genome/guide/human/>), we were able to elucidate the complex organization of the RAR β 5 and RAR β 2 genes. BLAST search permitted us to align the first three exons of RAR β 5 to a 70560-bp BAC clone RP11-421F9 (GenBank accession no. AC133141) mapped to chromosome 3p24. Similarly, the remaining exons were precisely aligned within another 189308-bp BAC clone RP11-659P16 (GenBank accession no. AC093416). Then the first exon of RAR β 2 was aligned within the 198468-bp BAC clone RP11-733H11 (GenBank accession no. AC098477). Additional alignment of these three clones showed a 4145-bp overlap between clones RP11-733H11 and RP11-421F9 and a 1862-bp overlap between clones RP11-421F9 and RP-659P16 (Fig. 8B), showing the continuity of the gene sequence in these BAC clones. An analysis of these three BAC clones revealed that the RAR β 5 gene spans over 130-kb of DNA, whereas the RAR β 2 gene spans over 160 kb of DNA. All of the splice junctions conform to the GT/AG rule for splice donor and acceptor sites (ref. 26; Fig. 8A). Fig. 8B summarizes our analysis on the genomic structures of hRAR β 5 and hRAR β 2 genes, and a new numeration for their exons is proposed herewith.

DISCUSSION

The major biological effect and gene expression induced by retinoids are believed to be mediated by nuclear receptors RARs/RXRs. Because RARs and RXRs are primary effectors of retinoid signaling, they themselves seem to be targets for disruption in tumorigenesis. RAR β has been extensively studied in human carcinomas, and several studies have suggested that it might play a role in tumor suppression (27–29). Therefore, RAR β has been considered a target molecule for retinoids in chemoprevention and therapeutic studies.

In this study, we identified RAR β 5, a novel RAR β isoform directed by a distinct promoter P3. We also provided the first evidence show-

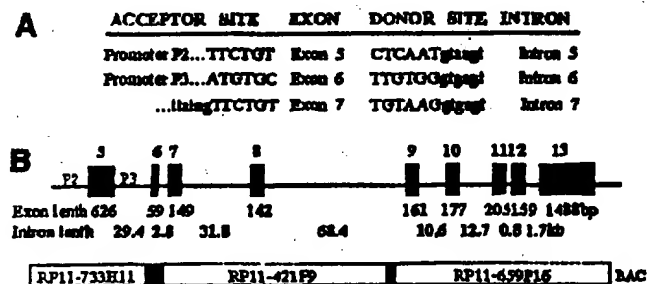


Fig. 8. Schematic representation of genomic structure of the hRAR β 5 gene in comparison with hRAR β 2. A, The intron-exon boundaries around exon 6 (the first exon of RAR β 5) are shown. Exon and intron sequences are represented by capital and lowercase letters, respectively. The canonical acceptor (ag) and donor (gt) splice sites are bolded. B, organization of the hRAR β 2 and hRAR β 5 genes and their alignment to BAC clones. RAR β 2 is driven by promoter P2, whereas RAR β 5 is directed by promoter P3. RAR β 2 and RAR β 5 only differ in their first exons; exons 7 to 13 are common to all of the RAR β isoforms. Exons are represented by black boxes, and lengths are shown in bp. Intron lengths (kb) were determined based on alignment of the cDNA sequence of RAR β 5 and RAR β 2 to BAC clones. The black boxes in BAC clones represent the overlapping region between the clones.

ing that RAR β 5 protein is identical to RAR β 4 or RAR β '³, a previously identified truncated RAR β protein (7, 11). In 1999, Sommer *et al.* (11) identified a 40 kDa RAR β protein isoform, which was interpreted as RAR β 4. This RAR β protein isoform was found to be elevated in human breast tumor cells, especially in cytoplasm relative to RAR β 2 protein. In 2002, Chen *et al.* (7) showed an antagonistic role for this RAR β protein isoform in signaling by retinoic acid and termed it RAR β ', and its expression was interpreted as "leaky scanning." In the same year, another group (15) showed that the expression of this protein isoform is associated with cellular resistance in response to retinoids. They also interpreted it as RAR β 4 and tried to link the protein expression to RAR β 4 mRNA expression. These data seemed correct, but the interpretation on the generation of that RAR β (RAR β 4 or RAR β ') protein isoform seems questionable. There has been no evidence showing that the protein isoform is translated from endogenous RAR β 2 or RAR β 4 transcripts. The interpretation on the generation of that RAR β protein isoform was based on transfection and *in vitro* translation experiments, in which the expression vectors generally do not contain full-length 5' and 3'UTR, and the reporter genes are not in a natural chromatin environment. In addition, the existence of multiple uORFs in the long 5'UTR region of RAR β 2 and RAR β 4 could also inhibit leaky scanning (30). Some cells such as MCF10A series of cell lines do not express detectable RAR β 4 mRNA,³ the same protein isoform could only be from RAR β 5 transcript in these cells. The identification of RAR β 5 in breast epithelial cells suggests that RAR β ' is the primary translation product from ORF of RAR β 5 transcript. Should leaky scanning occur with RAR β 2 or RAR β 4, it might result in RAR β ' at a very low level (30, 31). Moreover, the existence of multiple uORFs and the long leader sequence in RAR β isoform mRNAs could be a signal that their translations are under tight control.

In most of the previous studies, measurements of RAR β expression are preferentially made at the mRNA levels (25), leading to certain level of complexity in understanding its function. All the more, in most of breast cancer cell lines, RAR β 2 protein was not detected by Western blotting, although its mRNA was detectable (Fig. 2). We have carried out experiments to address these concerns to a certain extent and to study the expression of RAR β 5 both at the RNA and protein levels in various patient-derived primary breast cancer cells, established breast cancer cell lines, and immortalized benign MCF10A, premalignant MCF10AT cell lines, and normal human breast epithelial cells. At the mRNA level, RAR β 5 is expressed in normal human breast epithelial cells as well as in benign, premalignant, and tumor cell lines. In the presence of *atRA*, the level of RAR β 2 mRNA is preferentially elevated in contrast to RAR β 5 in T47D cells. At the protein level, we failed to detect endogenous RAR β 2 protein by Western blotting but did detect a corresponding RAR β 5 band in MDA-MB231 and HMEC cells. In agreement with our studies, Tanaka *et al.* (25) also failed to detect RAR β 2 protein under their experimental conditions. Hence, either RAR β 2 protein is not stable, or its expression is too low to be detected in these cells.

RAR β 5 identification also defines a new type of RAR β isoform, which is under the control of a distinct promoter P3, and the protein lacks the A, B, and part of the C domain (the first zinc finger) of other RAR β isoforms. The loss of DNA binding ability while retaining the capability to form heterodimers with RXR makes RAR β 5 act as a *trans*-dominant-negative regulator of RAR β function (7). We note in this respect that similarly truncated RAR α isoforms lacking all of the sequence located NH₂-terminal to the second zinc finger have been identified previously (32). Most analogous to RAR β 5 is the proges-

terone C mRNA that encodes a NH₂-terminally truncated progesterone receptor (33, 34). RAR β 5 protein expression, localization, and function were characterized previously as a truncated RAR β protein isoform (7, 11, 15). Unlike some other reported dominant-negative nuclear receptors (35, 36), RAR β 5 does not bind *cis*-acting DNA elements and therefore cannot directly inactivate gene transcription. RAR β 5 likely represses by stoichiometric competition, away from the RARE, against other transcription factors within the cell (e.g., RAR α , RAR β , and RAR γ) for transcription cofactors (7). Although RAR β 4 protein (identical to RAR β 5 protein) was reported to be elevated in breast cancer cells (11, 15), our data show that both RAR β 5 mRNA and protein are expressed in normal HMEC cells, indicating that RAR β 5 is not a tumor-specific isoform; it could be a regulatory factor for RAR β target genes in both normal and tumor breast epithelial cells. We could not detect RAR β 5 protein in ER-positive breast cell lines that are sensitive to retinoids, whereas it can be detected in ER-negative breast cancer cells and normal breast epithelial cells that are relatively resistant to retinoids, indicating that this isoform may contribute to cellular resistance to retinoids. In the metastatic *atRA*-resistant M-4A4 cell line derived from MDA-MB435 cells, RAR β 4 protein (identical to RAR β 5) was elevated in comparison with the isogenic nonmetastatic NM-2C3 cell line, and its protein expression was also up-regulated by *atRA* after 4- and 6-day treatment (15), which is consistent with our RT-PCR and MTT data and in agreement with the conclusion that it plays a negative role in RAR β 2 function (7).

Analysis of the RAR β 5 5' flanking region by computer program failed to predict the P3 promoter, indicating that P3 is not a typical promoter. The functionality of the TATA-like box 42-bp upstream of transcription start site remains unclear. In this respect, another non-canonical TATA box (TATATTA) has been reported in the P2 promoter of RAR β (9). However, cloning and transfection studies of the RAR β 5 5' flanking region confirmed the presence of P3 promoter. The *atRA* target promoter region (-302/-99) lacks any canonical RXRE/RARE elements. The magnitude of activation of the RAR β 5 promoter by *atRA* seemed to be cell-type specific. Cotransfection of empty vector pcDNA3.1 resulted in a significant increase in reporter gene activity, the reason is currently unknown; transfection experiment can generate artifacts, and a control (empty) vector must be included for comparison to see the function of the transfected gene. The effect of *atRA* seems to be at least partially RAR β 2 dependent, whereas RAR β 2 overexpression itself might not have a significant effect on RAR β 5 promoter activity in the absence of ligand-*atRA*. Other RAR/RXR selective retinoids also differentially increased RAR β 5 promoter activity, indicating that both RARs and RXRs can be involved in the RAR β 5 transcriptional activation. Because only two promoters have been previously identified in all of the RAR genes (11), the identification of this promoter has biological significance. Both RAR β 2 and RAR β 5 can be transactivated by *atRA* in the same cells, whereas their functions seem to be different, revealing a mechanism of fine-tuning *atRA*-induced transcription.

The corresponding isoform of RAR β 5 in RAR α and RAR γ gene seems not to be present, although fragments cleaved from RAR α and RAR γ protein were detected. Whether the fragments are functional or not and their biochemical properties still need to be determined. It seems RAR β 5 isoform is unique among all of the RARs and not a cleaved product from other RAR β isoforms, which suggest that RAR β gene might have different function from other RARs. The expression and regulation of RAR β protein is an important issue for functional research of this receptor. We did not observe significant regulation of RAR β 5 protein by *atRA* after 24 hours treatment with or without proteasomal inhibition, suggesting that the translational regulation is independent of transcriptional regulation under the experi-

³ X. Peng, R. G. Mehta, D. A. Tunetti, K. Christov, unpublished data.

mental condition. The RAR β post-translational regulation by retinoids will be addressed in depth in the future studies.

In summary, we have identified a novel, unique RAR β isoform (RAR β 5) and mapped its promoter region. We also initially characterized its expression and transcriptional regulation in normal and cancerous breast epithelial cells. RAR β 5 identification reveals an additional layer of complexity to retinoid signaling, and this isoform may serve as a potential target of retinoids in breast cancer prevention and therapy studies. Future study on RAR β function should include the analysis of RAR β 5 isoform in both normal and tumor cells and its response to retinoids. Effective inhibition of RAR β 5 might be necessary for the prevention and treatment of breast and other cancers by retinoids.

ACKNOWLEDGMENTS

We thank Dr. Karen Swisshelm for generously providing RAR β 2 expression plasmid, Dr. Zhenyu Li (Department of Pharmacology, University of Illinois at Chicago, Chicago, IL) for expert technical assistance in making DNA constructs, Dr. Koichi Shudo (ITSUU Laboratory) for providing Am80 RAR β selective retinoid, Marcia Dawson (Burnham Institute, La Jolla, CA) for providing RAR β agonists, and Scott Keundy for editorial assistance.

REFERENCES

- Giguere V. Retinoic acid receptors and cellular retinoid binding proteins: complex interplay in retinoid signaling. *Endocr Rev* 1994;15:61-79.
- Chambon P. A decade of molecular biology of retinoic acid receptors. *FASEB J* 1996;10:940-54.
- Gudas LJ, Sporn MB, Roberts AB. Cellular biology and biochemistry of the retinoids. In: Sporn MB, Roberts AB, Goodman DS, editors. *The retinoids*. 2nd ed. New York: Raven Press; 1994. p. 443-520.
- Dorand B, Saunders M, Lerry P, Leid M, Chambon P. All-trans and 9-cis retinoic acid induction of CRABP II transcription is mediated by RAR-RXR heterodimers bound to DR1 and DR2 repeated motifs. *Cell* 1992;71:73-85.
- Zhang XK, Lehmann J, Hoffmann B, et al. Homodimer formation of retinoid X receptor induced by 9-cis retinoic acid. *Nature (Lond)* 1992;358:581-91.
- Imeson K, Munkam K, Thompson CC, Evans RM. Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. *Cell* 1991;65:1235-66.
- Chen LJ, Sommer K, Swisshelm K. Downstream codons in the retinoic acid receptor β -2 and β -4 mRNAs initiate translation of a protein isoform that disrupts retinoid-activated transcription. *J Biol Chem* 2002;277:35411-21.
- de The H, Marchio A, Tiollais P, Dejean A. A novel steroid thyroid hormone receptor-related gene inappropriately expressed in human hepatocellular carcinoma. *Nature (Lond)* 1987;330:667-70.
- de The H, Vivanco-Ruiz MDM, Tiollais P. Identification of a retinoic acid response element in the retinoic acid receptor β gene. *Nature (Lond)* 1990;343:177-80.
- Nagpal S, Zelent A, Chambon P. RAR- β 4, a retinoic acid receptor isoform in generated from RAR- β 2 by alternative splicing and usage of a CUG initiator codon. *Proc Natl Acad Sci USA* 1992;89:2718-22.
- Sommer KM, Chen LJ, Treuting PM, Smith LT, Swisshelm K. Elevated retinoic acid receptor β 4 protein in human breast tumor cells with nuclear and cytoplasmic localization. *Proc Natl Acad Sci USA* 1999;96:8651-6.
- Houle B, Pelletier M, Wu J, Gourdyer C, Bradley WE. Fetal isoform of human retinoic acid receptor beta expressed in small cell lung cancer lines. *Cancer Res* 1994;54:365-9.
- Santocr SI, Dawson PJ, Teit L, et al. Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res Treat* 2001;65:101-10.
- Peng X, Yun D, Christov K. Breast cancer progression in MCF10A series of cell lines is associated with alterations in retinoic acid and retinoid X receptors and with differential response to retinoids. *Int J Oncology* 2004;25:961-71.
- Hayashi K, Goodison S, Urquidí V, et al. Differential effects of retinoic acid on the growth of isogenic metastatic and non-metastatic breast cancer cell lines and their association with distinct expression of retinoic acid receptor β isoforms 2 and 4. *Int J Oncol* 2003;22:623-9.
- Mehra RR, Birtescu L, Graves JM, et al. Human breast carcinoma cell lines: ultrastructural, genotypic, and immunocytochemical characterization. *Anticancer Res* 1992;12:583-92.
- Peng X, Maruo T, Matsuo H, Takekida S, Deguchi J. Serum deprivation-induced apoptosis in cultured porcine granulosa cells is characterized by increased expression of p53 protein, Fas antigen and Fas ligand and by decreased expression of PCNA. *Endocr J* 1998;45:247-53.
- Carmichael J, DeGruff WG, Gazdar AF, Minna JD, Mitchell JB. Evaluation of a tetrazolium-based semiautomated colorimetric assay: assessment of chemosensitivity testing. *Cancer Res* 1987;47:936-42.
- Kozak M. The scanning model for translation: an update. *J Cell Biol* 1989;108:229-41.
- Clifford JL, Menter DG, Wang M, Lotan R, Lippman SM. Retinoid receptor-dependent and -independent effects of N-(4-hydroxyphenyl)retinamide in F9 embryonal carcinoma cells. *Cancer Res* 1999;59:14-8.
- Prestridge DS. Predicting Pol II promoter sequence using transcription factor binding sites. *J Mol Biol* 1995;249:923-32.
- Knudsen S. Promoter 2.0: for the recognition of Pol II promoter sequences. *Bioinformatics* 1999;15:336-61.
- Quandt K, Frech K, Karas H, Wingender E, Werner T, Matted and Mathiaspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 1995;23:4878-84.
- Messeguer X, Escudero R, Farré D, Nuñez O. PROMO: detection of known transcription regulatory elements using species-biased searches. *Bioinformatics* 2002;18:333-4.
- Tanaka T, Rodriguez de la Concepcion ML, De Luca LM. Involvement of all-trans-retinoic acid in the breakdown of retinoic acid receptors alpha and gamma through proteasomes in MCF-7 human breast cancer cells. *Biochem Pharmacol* 2001;61:1347-55.
- Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* 1981;50:349-83.
- Deng G, Lu Y, Zlotnikov G, Thor AD, Smith HS. Loss of heterozygosity in normal tissue adjacent to breast carcinomas. *Science (Wash DC)* 1996;274:2037-9.
- Yang Q, Sakurai T, Kakudo K. Retinoic acid receptor β and breast cancer. *Breast Cancer Res Treat* 2002;76:167-73.
- Picard E, Seguin C, Monhoven N, et al. Expression of retinoic acid receptor genes and proteins in non-small-cell lung cancer. *J Natl Cancer Inst (Bethesda)* 1999;91:1059-66.
- Kozak M. Effects of long 5' leader sequences on initiation by eukaryotic ribosomes in vitro. *Gene Expr* 1991;1:117-25.
- Kozak M. An analysis of vertebrate mRNA sequences: intimations of translational control. *J Cell Biol* 1991;115:887-903.
- Lerry P, Krust A, Zelent A, et al. Multiple isoforms of the mouse retinoic acid receptor β are generated by alternative splicing and differential induction by retinoic acid. *EMBO J* 1991;10:59-69.
- Wei LL, Norris BM, Baker CJ. An N-terminally truncated third progesterone receptor protein, PR(C), forms heterodimers with PR(B) but interferes in PR(B)-DNA binding. *J Steroid Biochem Mol Biol* 1997;62:287-97.
- Wei LL, Hawkins P, Baker C, et al. An amino-terminal truncated progesterone receptor isoform, PRC, enhances progesterin-induced transcriptional activity. *Mol Endocrinol* 1996;10:1379-87.
- Pratt MA, Kralova J, McBurney MW. A dominant negative mutation of the alpha retinoic acid receptor gene in a retinoic acid-nonresponsive embryonal carcinoma cell. *Mol Cell Biol* 1990;10:6445-53.
- Dorand B, Saunders M, Gaudon C, et al. Activation function 2 (AF-2) of retinoic acid receptor and 9-cis retinoic acid receptor: presence of a conserved autonomous constitutive activating domain and influence of the nature of the response element on AF-2 activity. *EMBO J* 1994;13:5370-82.

ANTICANCER RESEARCH 24: 3965-3970 (2004)

Polymorphisms of Glutathione-S-Transferase and Arylamine N-Acetyltransferase Enzymes and Susceptibility to Colorectal Cancer

ISTVÁN KISS¹, ÁRPÁD NÉMETHI¹, BARNA BOGNER², GÁBOR PAJKOS³, ZSUZSA ORSÓS¹, JÁNOS SÁNDOR¹, ANDRÁS CSEJTEY⁴, ZSOLT FALUHELYI⁵, IMRE RODLER⁶ and ISTVÁN EMBER¹

¹Department of Public Health, Faculty of Medicine, Pécs University of Sciences, Pécs;

²Department of Pathology, Baranya County Hospital, Pécs;

³Central Hospital of Ministry of Internal Affairs, Budapest;

⁴Department of Oncoradiology, "Markusovszky" Vas County Hospital, Szombathely;

⁵Department of Oncology, Baranya County Hospital, Pécs;

⁶National Institute of Diet and Nutrition, Budapest, Hungary

Abstract. Background: Glutathione-S-transferases (GSTs) and N-acetyltransferases (NATs) are involved in the metabolism of a wide range of carcinogenic chemicals. Allelic polymorphism of these enzymes is associated with variations in enzyme activity, hence it may affect the concentration of activated carcinogenic chemicals in the body. Previous studies suggest a possible cancer risk-modifying effect of these allelic polymorphisms, but the results are still controversial. We evaluated the effect of GSTM1, GSTT1, GSTP1, NAT1 and NAT2 enzymes on individual susceptibility to colorectal cancer, with particular attention to possible interactions between the studied genotypes. Materials and Methods: Five hundred colorectal cancer patients and 500 matched cancer-free controls were included in the study. The allelic polymorphisms of GSTM1, GSTT1 and GSTP1, NAT1 and NAT2 enzymes were determined by PCR-based methods, from peripheral blood leukocytes, and allelic distributions were compared between colorectal cancer patients and controls. Results: The GSTM1 0 allele (OR: 1.48, 95% CI: 1.15-1.92) and rapid acetylator genotypes of NAT2 (OR: 1.52, 95% CI: 1.17-1.98) were associated with an elevated risk. No statistically significant correlation between NAT1, GSTT1, GSTP1 genotypes and colorectal cancer was found. Remarkably increased risk was associated with the GSTM1 0 allele - NAT2 rapid acetylator genotype combination (OR: 2.39, 95% CI: 1.75-3.26) and with the GSTM1 0 allele - NAT2 and

NAT1 rapid acetylator triple combination (OR: 3.28, 95% CI: 2.06-5.23). Carrying 4 or 5 putative "high-risk" alleles substantially increased the risk of colorectal cancer (OR: 3.69, 95% CI: 2.33-5.86). Conclusion: The genotype of certain metabolizing enzymes affects the risk for colorectal cancer. This effect is particularly important when certain allelic combinations are studied. In the near future, individual level risk assessment may be reached by further increasing the number of studied polymorphisms, combining them with traditional epidemiological risk factors.

It is generally accepted that cancer risk is determined by the interaction of environmental and genetic factors. Except for hereditary tumors, external carcinogenic exposure is involved in human tumorigenesis. Carcinogenic chemicals, however, undergo a complicated process of metabolism in the human body. Typically, these chemicals are activated by the so-called phase I metabolizing enzymes, which results in the formation of electrophilic, reactive compounds (1). The amount of active carcinogens is in good correlation with the risk of DNA damage and cancer formation. Detoxifying enzymes - phase II enzymes - help in the removal of carcinogens from the body (2). Most of these enzymes conjugate the carcinogenic chemical with a small molecule, making it less toxic and more water soluble. Therefore, it seems to be a logical assumption that the detoxifying capacity to a certain extent determines the individual susceptibility to cancer.

The activity of detoxifying enzymes in humans is basically determined by the genotype of the enzyme (2). Most of our metabolizing enzymes are genetically polymorphic, encoding proteins with different activities (2). Among the phase II enzymes, the glutathione-S-transferase (GST) superfamily and the N-acetyltransferases (NATs) have long been suspected to have an influence on cancer susceptibility (3-9).

Correspondence to: István Kiss, Department of Public Health, Faculty of Medicine, Pécs University of Sciences, Szégeti Str. 12, H-7643 Pécs, Hungary. Tel: (36) 72 536 395, Fax: (36) 72 536 395, e-mail: istvan.kiss@aok.pte.hu

Key Words: Colorectal cancer, metabolizing enzymes, cancer susceptibility, N-acetyltransferase, glutathione-S-transferase.

ANTICANCER RESEARCH 24: 3965-3970 (2004)

Table I. Allelic distributions of the studied GST and NAT enzymes in the control group.

	+	0	Ile/Ile	Heterozygous	Val/Val	Slow	Rapid
GSTM1	258	242	-	-	-	-	-
GSTT1	392	108	-	-	-	-	-
GSTP1	-	-	214	212	74	-	-
NAT1	-	-	-	-	-	305	195
NAT2	-	-	-	-	-	318	182

Table II. Allelic distributions of the studied GST and NAT enzymes among colorectal cancer patients.

	+	0	Ile/Ile	Heterozygous	Val/Val	Slow	Rapid
GSTM1	209	291	-	-	-	-	-
GSTT1	369	131	-	-	-	-	-
GSTP1	-	-	200	212	88	-	-
NAT1	-	-	-	-	-	289	211
NAT2	-	-	-	-	-	267	233

The GST enzymes have a relatively wide range of substrates, e.g. polycyclic aromatic hydrocarbons, monohalogenated, ethylene oxide, different solvents, pesticides (10). The superfamily consists of 6 families: α , μ , π , σ , θ and ξ . Probably, from a carcinogenic point of view, GSTM1, GSTT1 and GSTP1 are the most important enzymes, from the μ , θ and π families, respectively. In Caucasian populations, almost half of the individuals have no functional GSTM1 enzyme, due to a homozygous deletion in the gene (0 genotype) (11, 12). The situation is similar in the case of the T1 enzyme, but the ratio of persons with 0 genotype is lower (13). The GSTP1 enzyme possesses two single base polymorphisms, both resulting in an amino acid change in the protein (Ile105Val, Ala114Val) (14). In the case of the more frequently studied Ile105Val polymorphism, the Val allele encoded enzyme exhibits lower activity and, in accordance with this finding, certain tumors (e.g. lung, bladder) appeared to occur at higher rates among the carriers of the Val allele than among persons with the Ile genotype (15, 16).

The N-acetyltransferases are able to catalyze N- and O-acetylation, the former considered to be a detoxifying and the latter an activating reaction (17). Among their substrates, known carcinogenic compounds - like aromatic and heterocyclic amines - can be found (17). In the NAT family, polymorphisms of NAT2 are well characterized, but the NAT1 enzyme has only been recently studied from this point of view. Both NATs have several alleles; in the case of NAT2, the association of genotypes with enzyme activity is also well-established (usually people are categorized as rapid or slow acetylators) (18). The relationship between NAT1 alleles and acetylation speed is not so clear, but certain alleles also seem to be associated with the phenotype (19, 20).

Previous studies tried to find an association between the risk of different cancer types and the allelic polymorphism of GST and NAT enzymes. Regarding colorectal tumors, most of the studies suggested an elevated risk for individuals with the GSTM1 0 genotype (21-24). Based on theoretical considerations (because of O-acetylation of heterocyclic amines present in the GI system), rapid acetylators should also be at higher risk, but the results are controversial (25-28).

In the present case-control study, we tried to characterize the role of GSTM1, GSTT1, GSTP1, NAT1 and NAT2

polymorphisms in determining susceptibility to colorectal cancer. Since carriers of 0 alleles for the GST enzymes have a decreased detoxifying capacity, if this is combined with the rapid formation of metabolites of heterocyclic amines ensured by being a rapid acetylator, individuals with certain allelic combinations might be at a particularly high risk. Earlier, we demonstrated similar interactions between cytochrome P450 1A1 (CYP 1A1), cytochrome P450 2E1 (CYP 2E1) and GSTM1 alleles (29). The most important goal of the present study was to find such allelic combinations, and quantitatively assess their effect on colorectal cancer risk.

Materials and Methods

Five hundred colorectal cancer patients from the Central Hospital of the Ministry of Internal Affairs and from the area of Baranya and Vas County, Hungary, were included in the study. The diagnosis of tumors was always confirmed histologically. Patients with conditions affecting colorectal cancer risk (familial adenomatous polyposis, hereditary non-polyposis colorectal cancer, ulcerative colitis, etc.) were excluded from the study. Five hundred cancer-free controls from the same regions (non cancer patients from in- or outpatient wards and volunteers for health status examination) were matched to the cases according to age, sex, smoking habits, and red meat consumption. Ten ml peripheral blood was drawn from the participants, white blood cells were isolated by repeated centrifugation with 0.84% ammonium chloride and DNA was isolated (30).

GSTM1 and GSTT1 genotyping (31) was performed by a simultaneous amplification in the presence of an internal control (a 268 base length fragment of β -globin gene), with the following primers: GSTM1-F: GAACTCCCTGAAAAGCTAAAGC, GSTM1-R: GTTGGGCTCAAAATATACGGTGG, GSTT1-F: TTCCTTACTGGTCTCAGATCTC, GSTT1-R: TCACCGATC ATGGCCAGCA, β -globin-F: CAACITCATCCAGGTTTACC, β -globin-R: GAAGAGCCAAGGACAGGTAC. The reaction was performed in 20 μ l volume: 1.5 mM MgCl₂, 10 mM Tris-HCl (pH=8.3), 2 mg/ml bovine serum albumin, 4 x 0.25 mM dNTP, 2 U Taq DNS-polymerase, 30-30 pmol GSTT1-F and GSTT1-R primers, 50-50 pmol GSTM1-F and GSTM1-R primers, 20-20 pmol β -globin-F and β -globin-R primers, 13 μ l DNS-templat. After a 7-min denaturation at 94°C, 35 PCR cycles were performed: 60 sec 94°C, 60 sec 60°C, 60 sec 72°C, followed by 5 min at 72°C.

For GSTP1 the Ile105Val polymorphism was determined by a PCR-RFLP (32). A 176-bp fragment was amplified, with the

Kiss et al: GST and NAT Polymorphisms - Risk of Colon Cancer

Table III. Risk of colorectal cancer by genotypes of GST and NAT enzymes.

	Odds ratio	95% confidence interval
GSTM1	1.48	1.15-1.92
GSTT1	1.29	0.95-1.74
GSTP1	1.11	0.85-1.43
NAT1	1.14	0.88-1.48
NAT2	1.52	1.17-1.98

Table IV. Putative "high-risk" alleles per person in the control and case groups.

	Controls	Cases
0 "high-risk" allele per person	31	24
1 "high-risk" allele per person	120	119
2 "high-risk" alleles per person	185	131
3 "high-risk" alleles per person	134	135
4 "high-risk" alleles per person	29	75
5 "high-risk" alleles per person	1	16

following primers: 5'-ACCCAGGGCTCTATGGGA-3' and 5'-TGAGGGCACAAGAAGCCCT-3'. The reaction was carried out in 30 µl total volume, containing 50 ng DNA template, 4x200 µM dNTP, 200 ng each primer, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM MgCl₂ and 1 U Taq DNA polymerase. Parameters of the PCR reactions were as follows: 10 min at 95°C, then 30 cycles of 30 sec at 94°C, 30 sec at 55°C, and 30 sec at 72°C, followed by a final extension step at 72°C for 10 min.

The NAT2 allelic polymorphism was studied by restriction fragment length polymorphism (33). First, a nested PCR was used to amplify a 547 bp fragment of the gene (outer primer set: 5'-AATTAGTCACACGAGGA-3' and 5'-GCAGAGTGATTCATGCTAGA-3', inner set: 5'-CTGGGCTCTGGAAGCTCCTC-3', 5'-TTGGGTGATACATACACAAGGG-3', 25 cycles of 30 sec 94°C, 30 sec 59°C, 45 sec 72°C with the outer set was followed by 35 cycles with the inner set with the same parameters). NAT1*2*4 (wild-type), NAT2*5, NAT2*6 and NAT2*7 alleles were identified by restriction endonuclease digestion with *Kpn*I, *Taq*I, *Dde*I and *Ban*II enzymes. Homozygous or any heterozygous carriers of the wild-type allele were characterized as slow acetylators.

NAT1 genotyping was also undertaken using a nested PCR-based RFLP (33), similarly to the NAT2 genotyping, with the following primers: outer: 5'-GATCAAGTTGTGAGAAGAAATCGG-3', 5'-CTAGCATAAATCACCATTTCGAAG-3', inner: 5'-GACTCTGAGTGAGGTAGAAAT-3', 5'-CCACAGGCCATCTTTAGAA, at the underlined base constructing an additional *Mbo*II restriction site at the amplification of NAT1*4 allele. NAT1*4, NAT1*10 and NAT1*11 alleles were identified by this method, without studying certain rare alleles like NAT1*3 or NAT1*14. The presence of NAT*10 or NAT*11 alleles indicated the rapid acetylators.

Statistical calculations were made by Epi Info 6 (CDC, Atlanta, USA) and SPSS PC+ software. Odds ratios and 95% confidence intervals were used to compare the occurrence of genotypes in the case and control groups. In the case of the GSTM1 and GSTT1 + genotype, at GSTP1 homozygous *Ile* genotype and, in the case of NAT enzymes, slow acetylator genotypes were considered as baseline risk category.

Results

The allelic distributions in the control and case groups are shown in Tables I and II, respectively. The found allelic frequencies in the control group were similar to those of other studies in Caucasian populations. As illustrated in Table III, GSTM1 and NAT2 allelic distributions showed

statistically significant differences between cases and controls. There were no statistically significant effects of GSTT1, GSTP1 and NAT1 allelic polymorphisms on colorectal cancer risk. Composing subgroups within the NAT1 and NAT2 slow or rapid acetylators by exact genotypes did not give any further statistically significant result (data not shown).

Analyzing the joint effect of allelic combinations, GSTM1 and NAT2 alleles seemed to substantially strengthen each others effect: in the control group there were only 83 people possessing both "high-risk" alleles, while among the cases we found 161 such persons (OR: 2.39, 95% CI: 1.75-3.26). The paired analysis of GSTT1-GSTP1, GSTT1-NAT1 and GSTP1-NAT1 was also performed, but none of these combinations resulted in a statistically significant difference between cases and controls (data not shown). From triple combinations, GSTM1-NAT2-NAT1 caused the most remarkable difference, with an OR of 3.28 (95% CI: 2.06-5.23) for the simultaneous presence of the three "high-risk" alleles, suggesting a further risk-increasing effect by the third allele.

Since the analysis of allelic combinations suggested a possible interaction between the studied polymorphisms, we constructed a table based on the number of putative "high-risk" alleles per person among cases and controls (Table IV). The table clearly shows that persons with several "high-risk" alleles are relatively frequent among cases, while the control group mainly contains persons with fewer "high-risk" genotypes. When comparing the number of individuals with 4 or 5 "high-risk" alleles between cases and controls, the result is significantly different (OR: 3.69, 95% CI: 2.33-5.86). Interestingly, participants with less than 2 "high-risk" alleles were not significantly protected from developing colorectal cancer (OR: 0.93, 95% CI: 0.70-1.23).

Discussion

In our matched case control study, we found that carrying GSTM1 0 alleles or being a rapid acetylator were associated with an elevated risk of colorectal cancer in the studied Hungarian population. Unfortunately, several studies in the

ANTICANCER RESEARCH 24: 3965-3970 (2004)

field are not really comparable with each other, because some of them are not matched studies and, when matching is applied, the used variables may differ from each other. Further discrepancies may be caused by the different study populations: the allelic distributions might substantially differ from each other, not only in the studied polymorphism, but also in other genes which may also modify the risk of colorectal tumors.

The described problems can be seen when looking at the previous studies exploring the role of GSTM1 as a cancer risk modifier. The picture is confusing, since some studies suggested an association between 0 genotype and risk increase (34, 35), while others did not find any correlation (13). Our study, with relatively high case numbers, supports the hypothesis that the GSTM1 0 genotype is a risk factor of colorectal cancer susceptibility. This is in accordance with the detoxifying role of GSTM1 in the metabolism of carcinogenic substances.

The effect of GSTT1 polymorphism was not statistically significant, although it was near to that level (OR: 1.29, 95% CI: 0.95-1.74). Such results always raise the question of whether an increased sample size would result in a statistically significant association. Unfortunately, the study of low penetrance genes in human populations is fairly difficult, since existing associations might not be identified because of the presence of several confounding factors and the heterogeneity of the study population. This emphasizes the role of comparing and meta-analysis of different studies. Concerning the effect of GSTT1, it is of interest that, in spite of being near to the level of statistical significance, no effect in double or triple combinations was found, while NAT1, with a weaker effect alone, was part of a triple combination (GSTM1 0 genotype - NAT2 rapid acetylators - NAT1 rapid acetylators) which was associated with substantially elevated risk. In spite of the negative results of our study for the total sample, GSTT1 might be a risk modifier in certain subpopulations with heavy exposure to carcinogenic substances which are substrates of the GSTT1 enzyme.

GSTP1 polymorphisms have not been considered to play an important role in human colorectal carcinogenesis, however, its allelic polymorphism is associated with differences in the activity of the encoded enzymes. Here, we must not forget about the recently explored role of GSTs in cell signaling pathways, independently of their glutathione-S-transferase activity (36). GSTP1 is involved in the regulation of the MAP kinase pathway, by forming a complex with the c-jun N-terminal kinase. In the process of human carcinogenesis, GST enzymes as intracellular regulator proteins have been studied as possible factors with an influence on response to cytostatic treatment. From the cancer risk or cancer prevention point of view, this side of the GSTs has not been studied. Neither do we know whether allelic polymorphisms of GSTs affect their function

as intracellular regulators. Answering these questions might give further help in the explanation of the population level effects of GST alleles as cancer risk modifiers.

While the GST enzymes are important detoxifiers of metabolites of polycyclic aromatic hydrocarbons, NATs are involved in the metabolism of aromatic and heterocyclic amines. Since these compounds are present in our diet or are formed during food preparation, and NATs are present in the colorectal mucosa, there is a possible mechanistic link to explain the role of NAT polymorphisms in human carcinogenesis.

Allelic polymorphism of the NAT2 enzyme has been known for a long time, first detected phenotypically, based on enzyme activity distribution in healthy subjects, and later these activity differences were bound to an allelic polymorphism (37). Since NAT2 activates heterocyclic amines, rapid acetylators might be at higher risk of colorectal cancer formation. In our study, NAT2 polymorphism proved to be the strongest factor to affect the colorectal cancer risk. During recent years, the role of NAT2 seemed to be clarified by the previously mentioned model, which was also supported by epidemiological and molecular epidemiological facts. Particular importance was attributed to NAT2 in individuals with high red meat and/or well-done meat consumption (38), since these heterocyclic amine-containing dietary constituents served as sources of carcinogenic exposure. Some studies, however, seem to confuse the picture; a meta-analysis of D'Errico *et al.* found the NAT2 polymorphism not to be a significant risk factor (OR: 1.03, 95% CI: 0.93-1.14) (39), while a recent study of Sachse *et al.* did not find an association between NAT2 alleles and colorectal tumorigenesis (OR: 0.82, 95% CI: 0.69-1.12) (40), though still maintaining the connection between red meat consumption and colorectal cancer.

NAT1 was originally believed to be monomorphic, because of the unimodal distribution of its activity in the studied populations. Recently, several alleles have been identified and enzyme activity variations were also demonstrated; however, the phenotypical variations (enzyme activity differences) were lower than those measured in the case of NAT2 (19-20) alleles. Some recent studies also tried to demonstrate an association between NAT1 alleles and cancer risk. The results are controversial. Some studies identified NAT1 variants as risk factors (mainly the NAT1*10 allele was studied) (41, 42), while others did not demonstrate any association at all (43, 44). Further confusion is caused by discrepancies in the genotype-phenotype relationships reported by different authors. The NAT*10 allele is generally considered to be associated with higher activity, but some results seem to contradict these findings (45). Similarly, the activity of the NAT*11 allele is questionable. These contradictory results might be caused by tissue-specific differences in the expression of NAT enzymes,

Kiss *et al.*: GST and NAT Polymorphisms - Risk of Colon Cancer

as suggested by Bruhn *et al.* (45). Since we also performed an allele-specific analysis in the case of NAT1 and NAT2, resulting in the same associations as with large categories (rapid and slow acetylators), misclassification error caused by erroneously putting a genotype into the "slow" or "rapid" acetylator groups can be ruled out in our study.

Probably the most important part of studying the effects of low penetrance genes is the analysis of possible interactions between the investigated alleles. This might bring us to individual level risk assessment by giving a more precise estimation of the risk. From a practical point of view, the question is whether we are able to find such genetic conditions (allelic combinations) which considerably increase the cancer risk of a person. In our study such conditions included a triple combination with an OR of 3.28. A simple but very effective method for risk estimation is the calculation of simultaneously carried "high-risk" alleles. This method has the advantage of taking every existing interaction into consideration, studying actions as they happen, without including further possibilities of errors by introducing complicated mathematical modeling.

Our results (Table IV) support the hypothesis that even those allelic polymorphisms which did not have a significant influence on the risk of colorectal tumors, in certain still unknown circumstances or in not yet determined interactions, also slightly contribute to the modulation of the final risk. In our study, we demonstrated a substantially elevated risk in carriers of 4 or 5 "high-risk" alleles (OR: 3.69), but this still did not reach the "level of intervention". The results, however, allow us to hope that genotyping several polymorphisms simultaneously, together with the analysis of known traditional epidemiological risk factors, will give us the opportunity, in the near future, to estimate the individual susceptibility to the most important cancer types, allowing application of individually-shaped preventive strategies, or working out screening programs for identification of "high-risk" individuals.

Acknowledgements

This work was supported by the "Bolyai János" grant of the Hungarian Academy of Sciences.

References

- Meyer UA: Overview of enzymes of drug metabolism. *J Pharmacokinetic Biopharm* 24: 449-459, 1996.
- Taninger M, Malacarne D, Izzotti A, Ugolini D and Parodi S: Drug metabolism polymorphisms as modulators of cancer susceptibility. *Mutat Res* 436: 227-261, 1999.
- Ryberg D, Skaug V, Hewer A, Phillips DH, Harries LW, Wolf CR, OGREID D, Ulvik A, Vu P and Haugen A: Genotypes of glutathione transferase M1 and P1 and their significance for lung DNA adduct levels and cancer risk. *Carcinogenesis* 18: 1285-1289, 1997.
- Nakachi K, Imai K, Hayashi S and Kawajiri K: Polymorphisms of the CYP1A1 and glutathione S-transferase genes associated with susceptibility to lung cancer in relation to cigarette dose in Japanese population. *Cancer Res* 53: 2994-2999, 1993.
- Deakin M, Elder J, Hendrickse C, Peckham D, Baldwin D, Pantin C, Wild N, Leopard P, Bell DA, Jones P, Duncan H, Brannigan K, Alderson J, Fryer AA and Strange RC: Glutathione S-transferase GSTT1 genotypes and susceptibility to cancer: studies of interactions with GSTM1 in lung, oral, gastric and colorectal cancers. *Carcinogenesis* 17: 881-884, 1996.
- Potter JD, Bigler J, Fostick L, Bostick RM, Kampman E, Chen C, Louis TA and Grambsch P: Colorectal adenomatous and hyperplastic polyps: smoking and N acetyltransferase 2 polymorphisms. *Cancer Epidemiol Biomarkers Prev* 8: 69-75, 1999.
- Lower GM Jr, Nilsson T, Nelson CE, Wolf H, Gamsky TB and Bryan GT: N-acetyltransferase phenotype and risk in urinary bladder cancer: approaches in molecular epidemiology. Preliminary results in Sweden and Denmark. *Environ Health Perspect* 29: 71-9, 1979.
- Risch A, Wallace DM, Bathors S and Sim E: Slow N acetylation genotype is a susceptibility factor in occupational and smoking related bladder cancer. *Hum Mol Genet* 4: 231-236, 1995.
- Gertig DM, Hankinson SE, Hough H, Spiegelman D, Colditz GA, Willett WC, Kelsey KT and Hunter DJ: N-acetyltransferase 2 genotypes, meat intake and breast cancer risk. *Int J Cancer* 80: 13-17, 1999.
- Ketterer B, Taylor J, Meyer D, Pemble S, Coles B, Chulin X and Spencer S: Some function of glutathione transferases. In: Structure and Function of Glutathione Transferases (Tew K, Mannervik B, Mantle TJ, Pickett TJ, Hayes JD, eds.). Boca Raton, Florida, CRC Press, 1993, pp 15-27.
- Hirvonen A, Husgafvel-Pursiainen K, Anttila S and Vainio H: The GSTM1 null genotype as a potential modifier for squamous cell carcinoma of the lung. *Carcinogenesis* 14: 1479-1481, 1993.
- Board PG: Biochemical genetics of glutathione S-transferase in man. *Am J Hum Genet* 33: 36-43, 1981.
- Welfare M, Monesola Adeokun A, Bassendine MF and Daly AK: Polymorphisms in GSTP1, GSTM1, and GSTT1 and susceptibility to colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 8: 289-292, 1999.
- Henderson CJ, McLaren AW, Moffat GJ, Bacon EJ and Wolf CR: π -class glutathione S-transferase: regulation and function. *Chem Biol Interact* 171-172: 69-82, 1998.
- Ryberg D, Skaug V, Hewer A, Phillips DH, Harries LW, Wolf CR, OGREID D, Ulvik A, Vu P and Haugen A: Genotypes of glutathione transferase M1 and P1 and their significance for lung DNA adduct levels and cancer risk. *Carcinogenesis* (Lond.) 18: 1285-1289, 1997.
- Harries LW, Stubbins MJ, Forman D, Howard GC and Wolf CR: Identification of genetic polymorphisms at the glutathione S-transferase locus and association with susceptibility to bladder, testicular and prostate cancer. *Carcinogenesis* (Lond.) 18: 641-644, 1997.
- Evans DA: N-acetyltransferase. In: Pharmacogenetics of Drug Metabolism (Kalow W ed). New York, Pergamon Press, 1992, pp 95-178.
- Kadlubar FF, Butler MA, Kadlulik KR, Chou HIC and Lang NP: Polymorphisms for aromatic amine metabolism in humans: relevance for human carcinogenesis. *Environ Health Perspect* 98: 69-74, 1992.

ANTICANCER RESEARCH 24: 3965-3970 (2004)

- 19 Kukongviriyapan V, Prawan A, Warasit B, Tassanayakul W and Aiamsa-ard J: Polymorphism of N-acetyltransferase 1 and correlation between genotype and phenotype in a Thai population. *Eur J Clin Pharmacol* 59: 277-81, 2003.
- 20 Loktionow A, Moore W, Spencer SP, Vorstic H, Nell T, O'Neill IK, Bingham SA and Cummings JH: Differences in N-acetylation genotypes between Caucasians and Black South Africans: implications for cancer prevention. *Cancer Detect Prev* 26: 15-22, 2002.
- 21 Welfare M, Montesola Adekun A, Bassendine MF and Daly AK: Polymorphisms in GSTP1, GSTM1, and GSTT1 and susceptibility to colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 8: 289-292, 1999.
- 22 Kato T, Nagata N, Kuroda Y, Itoh H, Kawahara A, Kuroki N, Ookuma R and Bell DA: Glutathione S transferase M1 (GSTM1) and T1 (GSTT1) genetic polymorphism and susceptibility to gastric and colorectal adenocarcinoma. *Carcinogenesis* 17: 1855-1859, 1996.
- 23 Kampman E, Slattery ML, Bigler J, Leppert M, Samowitz W, Caan BJ and Potter JD: Meat consumption, genetic susceptibility, and colon cancer risk: a United States multicenter case control study. *Cancer Epidemiol Biomarkers Prev* 8: 15-24, 1999.
- 24 Slattery ML, Potter JD, Ma KN, Caan BJ, Leppert M and Samowitz W: Western diet, family history of colorectal cancer, NAT2, GSTM1 and risk of colon cancer. *Cancer Causes Control* 11: 1-8, 2000.
- 25 Chi JP and Lechner MC: Increased frequency of wild type arylamine N acetyltransferase allele NAT2*4 homozygotes in Portuguese patients with colorectal cancer. *Carcinogenesis* 19: 37-41, 1998.
- 26 Ladero JM, Gonzalez JF, Benitez J, Vargas E, Fernandez MJ, Baki W and Diaz-Rubio M: Acetylation polymorphism in human colorectal carcinoma. *Cancer Res* 51: 2098-2100, 1991.
- 27 Spurr NK, Gough AC, Chingwundoh FI and Smith CAD: Polymorphisms in drug-metabolizing enzymes as modifiers of cancer risk. *Clin Chem* 41: 1864-1869, 1995.
- 28 Le Marchand L, Hankin JH, Wilkens LR, Pierce LM, Franke A, Kolonel LN, Seifried A, Custer LJ, Chang W, Lum-Jones A and Donlon T: Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev* 10: 1259-66, 2001.
- 29 Kiss I, Sándor J, Pajkos G, Bogner B and Ember I: Colorectal cancer risk in relation to genetic polymorphism of cytochrome P450 1A1, 2E1, and glutathione-S-transferase M1 enzymes. *Anticancer Res* 20: 519-522, 2000.
- 30 Blin N and Stafford DW: A general method for isolation of high molecular weight DNA from eucaryotes. *Nucleic Acids Res* 3: 2303-8, 1976.
- 31 Pool-Zobel BL, Bub A, Liegibel UM, Treptow van Lishaut S and Rechkemmer G: Mechanism by which vegetable consumption reduces genetic damage in humans. *Cancer Epid Biom Prev* 7: 891-899, 1998.
- 32 Jerónimo C, Varzim G, Henrique R, Oliveira J, Bento MJ, Silva C, Lopes C and Sidransky D: H105V Polymorphism and promoter methylation of the GSTP1 gene in prostate adenocarcinoma. *Cancer Epidemiol Biomarkers Prev* 11: 445-50, 2002.
- 33 Okkels H, Sigsgaard T, Wolf H and Autrup H: Arylamine N-acetyltransferase 1 (NAT1) and 2 (NAT2) polymorphisms in susceptibility to bladder cancer: the influence of smoking. *Cancer Epidemiol Biomarkers Prev* 6: 225-31, 1997.
- 34 Gawronska Szklarz B, Lubinski J, Kladny J, Kurzawski G, Bielicki D, Wojcicki M, Sych Z and Musial HD: Polymorphism of GSTM1 gene in patients with colorectal cancer and colonic polyps. *Exp Toxicol Pathol* 51: 321-325, 1999.
- 35 Zhong S, Wyllie AH, Barues D, Wolf CR and Spurr NK: Relationship between GSTM1 genetic polymorphism and susceptibility to bladder, breast and colon cancer. *Carcinogenesis* 14: 1821-1824, 1993.
- 36 Townsend DM and Tew KD: The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* 22: 7369-7375, 2003.
- 37 Le Marchand L, Sivaraman L, Franke AA, Custer LJ, Wilkens LR, Lau AF and Cooney RV: Predictors of N-acetyltransferase activity: should caffeine phenotyping and NAT2 genotyping be used interchangeably in epidemiological studies? *Cancer Epidemiol Biomarkers Prev* 5: 449-455, 1996.
- 38 Robert-Thomson IC, Ryan P, Khoo KK, Hart WJ, McMichael AJ and Butler RN: Diet, acetylator phenotype and risk of colorectal neoplasia. *Lancet* 347: 1372-1374, 1996.
- 39 D'Errico A, Malata N, Vincis P and Bufetta P: Review of studies of selected metabolic polymorphisms and cancer. In: *Metabolic Polymorphisms and Susceptibility to Cancer* (Vincis P et al. eds). IARC Sci. Publ. 148. IARC, Lyon, France. Pp 323-393.
- 40 Sachse C, Smith G, Wilkie MJV, Barrett JH, Waxman R, Sullivan F, Forman D, Bishop DT, Wolf CR and the Colorectal Study Group: A pharmacogenetic study to investigate the role of dietary carcinogens in the etiology of colorectal cancer. *Carcinogenesis* 23: 1839-1849, 2002.
- 41 Bell DA, Stephens E, Castranio T, Umhach DM, Watson M, Deakin M, Elder J, Duncan IL, Hendricks C and Strange RC: Polyadenylation polymorphism in the N-acetyltransferase gene 1 (NAT1) increases risk of colorectal cancer. *Cancer Res* 55: 3537-3542, 1995.
- 42 Chen J, Stampfer MJ, Hough HL, Garcia-Closas M, Willett WC, Hennekens CH, Kelsey KT and Hunter DJ: A prospective study of N-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. *Cancer Res* 58: 3307-3311, 1998.
- 43 Probst-Hensch NM, Haile RW, Li DW, Sakamoto GT, Louie AD, Lin BK, Frankl HD, Lee ER and Lin HJ: Lack of association between the polyadenylation polymorphism in the NAT1 (acetyltransferase 1) gene and colorectal adenomas. *Carcinogenesis (Lond.)* 17: 2125-2129, 1996.
- 44 Lin HJ, Probst-Hensch NM, Hughes NC, Sakamoto GT, Louie AD, Kau IH, Lin BK, Lee DB, Lin J, Frankl HD, Lee ER, Hardy S, Grant DM and Haile RW: Variants of N-acetyltransferase NAT1 and a case-control study of colorectal adenomas. *Pharmacogenetics* 8: 269-281, 1998.
- 45 Bruhn C, Brockmüller J, Casanovi I, Roots I and Borchert H-H: Correlation between genotype and phenotype of the human arylamine N-acetyltransferase type 1 (NAT1). *Biochem Pharmacol* 58: 1759-1764, 1999.

Received April 7, 2004

Revised September 8, 2004

Accepted November 14, 2004

N- and C-Terminal Isoforms of Arg Quantified by Real-Time PCR Are Specifically Expressed in Human Normal and Neoplastic Cells, in Neoplastic Cell Lines, and in HL-60 Cell Differentiation

Roberto A. Perego,^{1*} Matteo Corizzato,¹ Cristina Bianchi,¹ Barbara Eroini,¹ and Silvano Bosari²

¹Department of Experimental & Environmental Medicine and Medical Biotechnologies, School of Medicine, Milano-Bicocca University, Monza (MI), Italy

²Department of Medicine, Surgery and Dentistry, Pathology Unit, School of Medicine, Milano University, A.O. S. Paolo and IRCCS Ospedale Maggiore, Milan, Italy

The human ABL2 (or ARG) gene codes for a nonreceptor tyrosine kinase is involved in translocation with the ETV6 gene in human leukemia and has an altered expression in several human carcinomas. Two isoforms of Arg with different N-termini (1A and 1B) have been described. The C-terminal domain of Arg contains two F-actin-binding sequences that perform a number of actions related to cell morphology and motility by interacting with actin filaments. We have identified different-sized specific cDNAs in hematopoietic, epithelial, nervous, and fibroblastic cells by means of the reverse transcription (RT)-polymerase chain reaction (PCR) analysis of human Arg mRNA. Some of these cDNAs showed an adjunctive alternative splice event involving the 63 bp sequence of exon II, thus leading to four cDNA types with different N-termini: 1A long and short, and 1B long and short. Other cDNAs lacked a 309 bp sequence in the last exon involving one of the C-terminal F-actin binding domains, thus giving rise to two cDNA types: C-termini long and short. Quantified by real-time PCR—quantitative RT-PCR—these Arg transcript isoforms have specific expression patterns not only in different normal and tumor cell types, but also during cell differentiation and growth arrest. These isoforms maintained the open reading frames, and eight putative proteins were predicted. The different C-termini isoforms seem to retain the same quantitative reciprocal ratio of their respective transcripts. The Arg protein isoforms with different C-terminal actin-binding domains and different N-termini might have specific cellular localizations/concentrations, and differently regulated catalytic activity with different implications in normal and neoplastic cells. © 2005 Wiley-Liss, Inc.

Key words: Arg tyrosine kinase; mRNA splicing; transcript expression; protein expression; actin-binding sequence

INTRODUCTION

The Abelson family of nonreceptor tyrosine protein kinases is defined by products of human and mouse ABL2 (also known as ARG, Abelson Related Gene) and ABL1 genes, and Drosophila and nematode ABL genes [1–3]. In human acute leukemia, ABL2(ARG) gene can be involved in translocations with the ETV6 gene and produce different chimeric proteins [4–6]. An altered expression of Arg transcripts has been described in different tumors [7–10]. The human Arg protein has a high degree of amino acid sequence identity (90%–94%) with c-Abl in the tyrosine kinase SH2 and SH3 domains [3]. Two isoforms of both human Arg and c-Abl have been described as having different N-termini called 1A and 1B [3,11]. Four mouse c-Abl isoforms have been cloned with different 5'-ends arising as a result of the addition of alternative 5'-exons [12]. Although the long C-terminal domain of Arg is quite different from that of c-Abl, both contain three proline-rich

sequences that bind to the SH3 domains of adaptor proteins [13,14]; c-Abl contains only one F-actin-binding sequence, while Arg contains two plus one microtubule-binding sequences [1,15]. The Arg product is located in the cytoplasm [16] whereas the c-Abl one is also nuclear.

Abbreviations: RT-PCR, reverse transcription-polymerase chain reaction; 1AL, 1AS, 1BL, 1BS, long and short isoforms of the 5'-end (N-termini) type 1A and 1B of Arg transcript or protein; ATRA, *All-trans* retinoic acid; TPA, 12-O-tetradecanoyl-phorbol-13-acetate; GM-CSF, granulocyte-macrophage colony stimulating factor; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; CTL, CTS, long and short isoforms of the 3'-end (C-termini) of Arg transcript or protein.

*Correspondence to: Department of Experimental & Environmental Medicine and Medical Biotechnologies, Via Cadore 48, 20052 Monza (MI), Italy.

Received 3 June 2004; Accepted 16 December 2004

DOI 10.1002/mc.20085

Published online in Wiley InterScience (www.interscience.wiley.com)

The functional role of Arg is currently under investigation. Through its interactions with actin filaments, it performs redundant actions with c-Abl, playing a role in neurulation [17]. It has a function in adhesion-dependent neuritogenesis [18], and in synaptic structure and function [19,20]. Arg seems to be required for bacterial pathogenesis [21], and to be involved with c-abl in oxidative stress response [22] by regulating catalase activity [23]. The suppression of Arg kinase activity by STI571 induces cell cycle arrest [24], and it appears that Arg plays a role in homologous recombination DNA repair [25]. Lymphopenia occurs during the development of mice harboring a homozygous disruption of c-Abl [26], thus indicating that Arg is unable to substitute c-Abl functions in lymphoid tissues. Arg is ubiquitous with greatest expression in nervous tissues [27]. Arg mRNA increases during granulocytic and macrophage-like differentiation of HL-60 cells [28], and its expression is higher in mature than in immature B lymphoid cell lines [29].

During reverse transcriptase (RT)-polymerase chain reaction (PCR) analyses of human Arg mRNA, we had identified specific cDNAs of different sizes. These showed an adjunctive splice event immediately downstream of both the alternatively spliced 1A and 1B exons, and the lack of a sequence in the last exon coding the C-termini. These events gave rise to four cDNA types diverging at the 5'-end and two cDNA types diverging in the 3'-region. Their open reading frames were maintained, and the possible combinations of the different splicing events made it possible to predict eight putative proteins. There was a differential expression of the Arg transcript isoforms quantified by real-time PCR—quantitative RT-PCR—under diverse physiological conditions and in normal and tumor cells.

MATERIALS AND METHODS

Cells, Tissues, and Human Cell Lines

Unpooled samples of lymphocytes, monocytes and granulocytes were obtained from volunteer donors by means of density gradient separation in Ficoll-Hypaque and Percoll as described [30,31]. Purity (>90%) was determined microscopically after May Grunwald Giemsa staining. The leukemic blast cells were obtained at diagnosis from bone marrow of patients affected by acute myelogenous leukemia and separated by sedimentation on Ficoll-Paque gradient as mononuclear fraction. The leukemic blasts were >90% of total cells. These were characterized as myeloid (M1), monocytic (M5) blasts according to FAB classification [32]. Tumor tissue specimens (renal clear cell carcinoma, Grade 2; colon carcinoma, Dukes histopathological stage A, Grade 1) and corresponding normal tissues (renal cortex; colon mucosa) were obtained from patients soon after surgical treatment; fresh tissue fragments were

immediately put into RNAlater (Ambion, Austin, TX) and frozen down in liquid nitrogen.

The human cell lines used (Table 1) were cultured with RPMI 1640 medium supplemented with 10% fetal calf serum, and tested during exponential growth. The growth characteristics, the differentiation of HL-60 cells to granulocytes by means of 4-d treatment with 1 μ M all-*trans* retinoic acid (ATRA) and to macrophage-like cells by means of 2-d treatment with 10 nM 12-O-tetradecanoyl-phorbol-13-acetate (TPA), as well as the immunofluorescence analysis of membrane CD11b marker expression, were performed as previously described [28]. The GFD8 cell line was cultured in the presence or in the 4-d absence of 5 ng/ml of granulocyte-macrophage colony stimulating factor (GM-CSF) for which it displays growth dependence. Removal of growth factor for 4 d led to reversible growth arrest in the absence of differentiation [34,35]. The growth characteristics of GFD8 cells were assessed by daily count and expression of the CD11b differentiation marker [28]. The ATRA, TPA, and GM-CSF came from Sigma-Aldrich (St. Louis, MO).

RNA Extraction, cDNA Synthesis, and Qualitative RT-PCR Analysis

Total RNA was obtained by cell and tissue extraction with TRIZOL (Invitrogen, Carlsbad, CA)

Table 1. Characteristics of Human Cell Lines Utilized

B Lymphoid	
LP-1 (myeloma, mature plasma cell phenotype)	
Raji (lymphoma, mature B cell phenotype)	
ALLPO (acute leukemia, immature early pre B cell phenotype)	
T Lymphoid	
Jurkat (acute leukemia, mature post thymic phenotype)	
Molt-4 (acute leukemia, immature thymocyte phenotype)	
Myeloid	
K562 (chronic leukemia, erythroid lineage phenotype)	
HL-60 (acute leukemia, granulocytic lineage phenotype)	
GFD8 (acute leukemia, granulocytic lineage phenotype)	
U937 (histiocytic lymphoma, monocytic lineage phenotype)	
Neuronal	
A-172 (glioblastoma)	
Lan-5 (neuroblastoma)	
Epithelial	
Caki-1 (renal cell carcinoma, clear cells)	
Hela (cervix carcinoma)	
Fibroblastic	
Hel 299 (lung embryonic fibroblast)	

The cell lines were from American Type Culture Collection, except LP-1 and Lan-5 that were from German Collection of Microorganism and Cell Cultures, and ALLPO [33] and GFD8 [34] that were a kind gift from A. Biondi (Milano-Bicocca University, Monza, Italy).

according to the manufacturer's instruction; it was spectrophotometrically quantified and its integrity was analyzed by electrophoresis in 1% agarose gel. The DNase treatment of total RNA and the reverse transcription of an 8 µg aliquot of DNA-free RNA in a 40 µl reaction in the presence of 0.5 µg of random examers was performed as previously described [28]; 2.5 µl cDNA was amplified in the presence of 0.4 µM primers, 2 mM MgCl₂, 0.2 µM dNTP, 2.5 U Taq Gold polymerase and 1x manufacturer's buffer (Applied Biosystem, Foster City, CA). The primers used in the combinations described in Figures 1 and 2 had the following sequences [3] and localizations:

41N 5'-ACACAGGTCCATGGTACC-3' reverse (exon IV)
 42N 5'-GCAGAGATCAGGACACTT-3' sense (exon 1A)
 132N 5'-AAGCTCGGGGCTCCAGC-3' sense (exon 1B)
 112N 5'-CACCAGGGATAGGAAGGGG-3' sense (exon XII)
 113N 5'-GGGAAGGGTCATTGCCATC-3' reverse (exon XII)
 114N 5'-CTGCTCTGGAAGCCcggtg-3' reverse (exon XII)
 115N 5'-ACCAGATTCGCTCTTGCTG-3' reverse (exon XII)

41N, 42N, 132N primers have an additional 5' eight-nucleotide tail containing the *Eco*RI restriction site. The capital and lower case letters of primer 114N show the fusion point of the sequences that juxtapose after the loss of a 309 bp fragment in the 3'-end of Arg cDNA. The amplification program was 95°C/10 min, (94°C/30 s, 60°C/30 s, 72°C/30 s) × 40 cycles, and 72°C/10 min. All the amplified cDNA were sequenced with the ABI Prism Kit Big Dye Terminator v3.0 sequencing kit, and the ABI Prism 3100 Avant Genetic Analyzer. The intron-exon junction of ABL2(ARG) (Figure 1) was determined with the NCBI Genome Map Viewer *Homo sapiens* database, Build 34, Version 1 (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606&query=arg).

Quantitative Real-Time PCR Analysis

Real-time PCR with TaqMan chemistry was used to quantify specific cell mRNA. The amplification was performed in an ABI PRISM 7900HT Sequence Detector. One microliter of the RT reaction (corre-

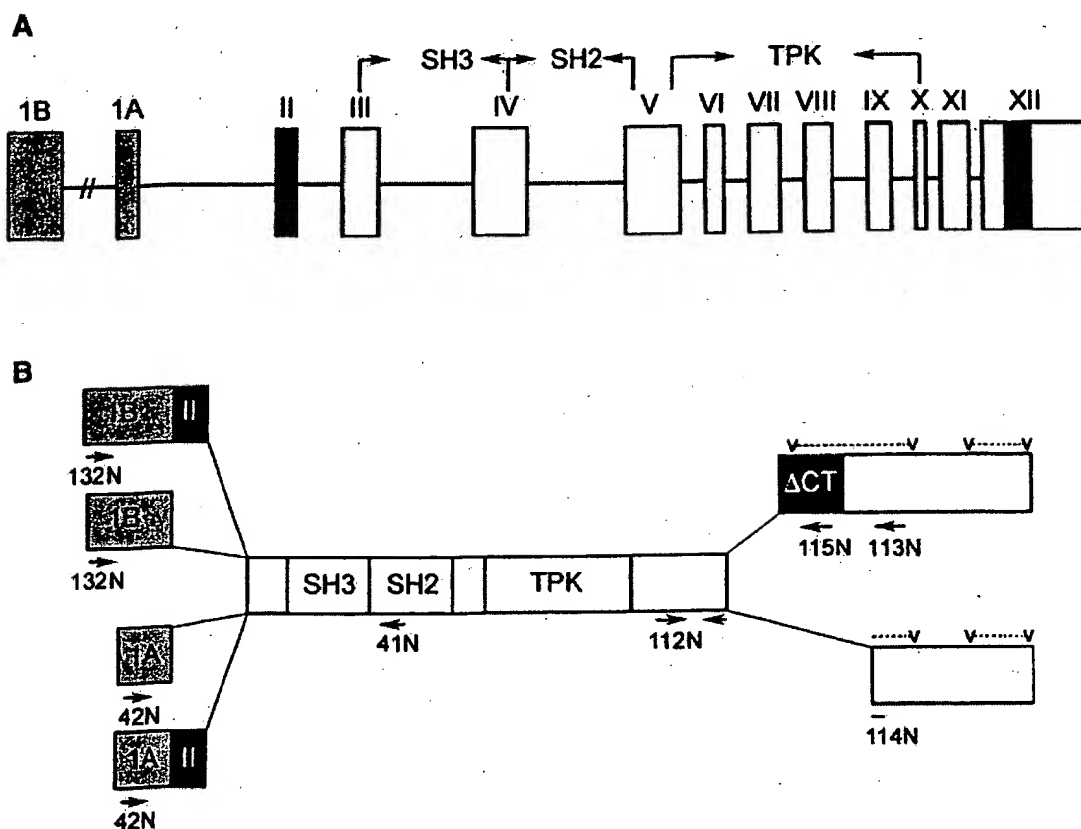


Figure 1. Schematic representation of the structure and cDNA coding sequences of the ABL2(ARG) gene. (A) The intron-exon structure of the ABL2(ARG) gene derived from the NCBI Genome Map View *Homo sapiens* database, Build 34, Version 1 and our own data. (B) Proposed isoforms of Arg proteins as predicted by the sequence of Arg cDNA obtained by RT-PCR analysis. The first alternative exons 1A and B are shaded, and the alternative spliced

exon II (II) and C-terminal region (ΔCT) lacking in the CTS form are shown in black. SH3-SH2 and TPK indicate the SH3, SH2, and tyrosine protein kinase domains. The sense and reverse primers used in the RT-PCR are indicated with arrows. Primer 114N spans noncontiguous sequences. The position of the two F-actin-binding regions in the C-terminal domain are also indicated (v...v). The diagrams are not to scale.

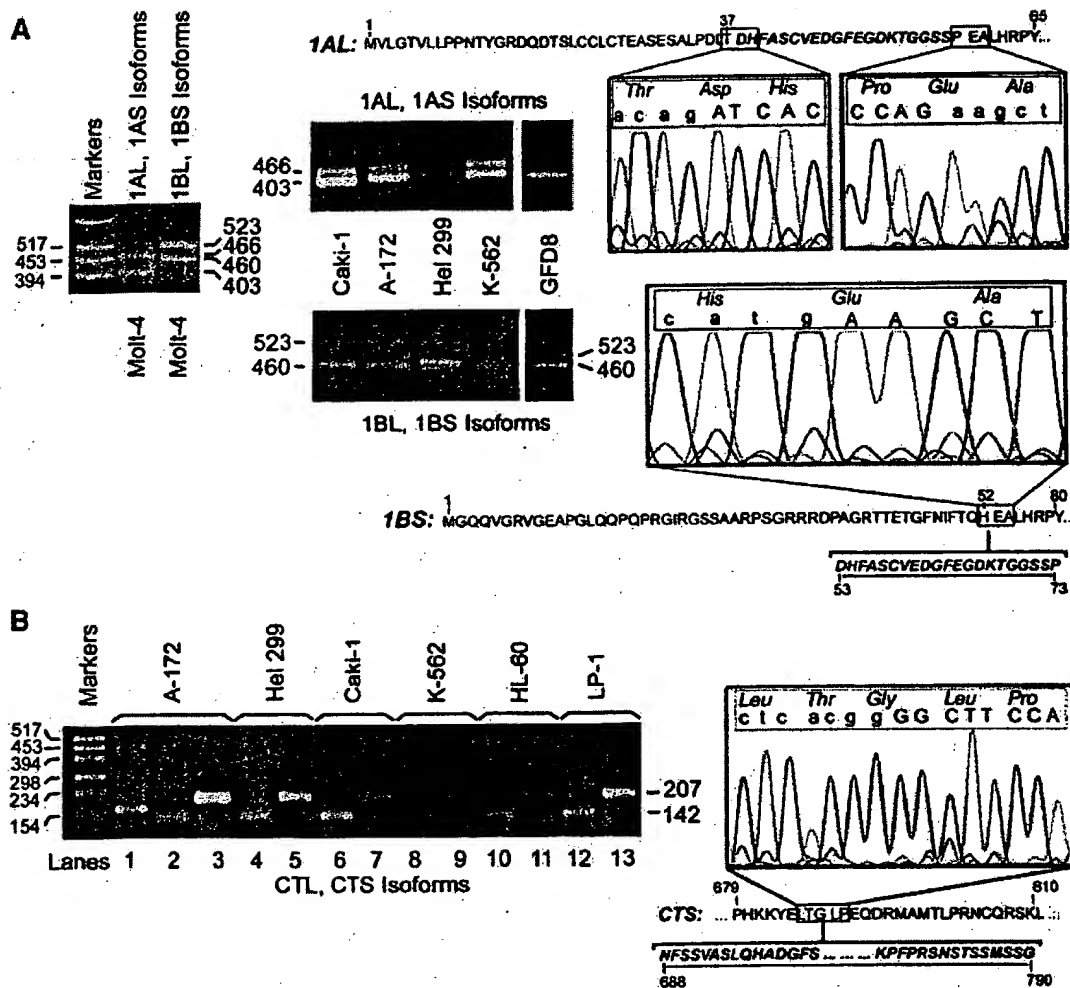


Figure 2. Detection of different ARG cDNA isoforms in various cell types by means of qualitative RT-PCR analysis. (A) 5'-end ARG isoforms. Left: The 41N/42N primer pair (see Figure 1) revealed two bands of 466 and 403 bp, corresponding to the 1AL and 1AS transcript isoforms, and the 41N/132N primer pair amplified two bands of 523 and 460 bp, corresponding to the 1BL and 1BS transcript isoforms. Right: Amino acid sequences of the 1AL (top) and 1BS isoform (bottom). The in-frame nucleotide sequence at the splicing sites (denoted by capital and lower case letters) are given. The 21 amino acids of exon II present in the 1AL form are indicated in bold. (B) 3'-end ARG isoforms. Left: The use of 112N/113N primer pairs (see Figure 1) led to obtaining two bands of 475 and 166 bp in

A172 cell lines (lane 1). In these cells and other cell types, mRNA amplification with the 112N/114N (even lanes) and 112N/115N primer pairs (odd lanes from lane 3), respectively, evidenced a 142 and a 207 bp specific band, corresponding to the CTS and CTL isoforms. Right: Amino acid sequence of the CTS isoform and the in-frame site of the nucleotide sequences that juxtapose (lower case and capital letters) after the loss of the 309 bp fragment. The 21 amino acid sequences lacking in the 1BS isoform and the 103 amino acid sequences lacking in the CTS isoforms are underlined; the reported amino acid numbers are those of the entire coding sequence.

responding to 200 ng of cDNA) was amplified in a 50 μ l PCR mixture, containing 1 \times Universal PCR master mix (Applied Biosystems) and different concentrations of primers and probes (Table 2) whose sequences were selected with Primer Express 2.0 software (Applied Biosystems). The transcript of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) housekeeping gene [36] was amplified as an endogenous control of RNA quality. Each cell line underwent at least two independent experiments, with

each sample being analyzed in triplicate. The real-time PCR conditions were 50°C/2 min (for optimal AmpErase uracil-N-glycosylase [UNG] activity), 95°C/10 min followed by (95°C/15 s and 60°C/1 min) \times 40 cycles. The ABI7900 system software raised the threshold cycle (C_T) values representing the cycle numbers needed to reveal the minimal amount of amplified material of a target transcript. The relative levels of the total Arg transcript in each sample were calculated with the averaged C_T values

Table 2. Sequences, Exon Localizations, and Concentrations of Primers and Probes Used for Transcript Quantification by Real-Time PCR and Their Respective Amplicon Lengths

Forms	Primer sense	Exon	nM	Primer reverse	Exon	nM	Probe ^a	Exon	nM	Amplicon (bp)
Arg Tot	(O) 5'AGTTTAGCACCAGgggtatca3' ^a	XI-XII	300	(P) 5'CTTCTATCTCCCTGGTGAAGCAT3'	XII	300	(V) 5'CACAGGCTCTAGTGGATCCCAGC3' (sense)	XII	200	135
Arg 1A	(C) 5'TGCTCTACCCGACTTAAACAGatc3'	1A-II	900	(B) 5'aaggttCTGGACTACTGCTCC3'	II-III	900	(K) 5'ATGCTGTGGAGGTGGATTTGAGGG3' (sense)	II	150	92
Arg 1AS	(F) 5'TGCTCTACCCGACTTAAACAGatc3'	1A-III	300	(E) 5'ACCTGATAGCTTATTTAGTCC3'	III	300	(J) 5'ATGCTGTGGAGGTGGATTTGAGGG3' (sense)	III	150	83
Arg 1BL	(A) 5'AATATCTTACCCAGCATGatca3'	1B-II	300	(B) 5'aaggttCTGGACTACTGCTCC3'	II-III	900	(K) 5'ATGCTGTGGAGGTGGATTTGAGGG3' (sense)	II	150	89
Arg 1BS	(D) 5'CAATCTTACCCAGCATGatca3'	1B-III	300	(E) 5'ACCTGATAGCTTATTTAGTCC3'	III	900	(J) 5'ATGCTGTGGAGGTGGATTTGAGGG3' (sense)	II	150	82
Arg CTL	(S) 5'CCTTCCGAGAAATGGAGATCA3'	XII	300	(T) 5'GCATGCTGTAGAGAAACAGAG3'	XII	300	(V) 5'CCCCATAAGAAATACGAACTCAGGGTAACTTCT3' (sense)	XII	200	82
Arg CTS	(Q) 5'TGGAGAAATCAGCCCAATAGAG3'	XII	300	(R) 5'TCTTGGGAAAGGTCATTGC3'	XII	300	(V) 5'ATACGAACTCAGGGgttccagag3' (sense)	XII	200	78
GAPDH	(L) 5' GAAGGTGAAGGTCGGAGTC3'	II	200	(M) 5'GAAGATGGTGGATTC3'	IV	200	(W) 5'CAAGCTTCCCTCTCAGCC3' (reverse)	IV	100	226

^aThe capital and lower case letters show the sequences located on different exons and, in the case of probe (V), the sequences that juxtapose after the loss of 309 bp fragment in exon XII of Arg.

^bThe TaqMan probes were labeled at the 5'-end with the reporter dye molecule FAM (6-carboxy-fluorescein) (Arg probes) or VIC (GAPDH probe), and at the 3'-end with the quencher dye molecule TAMRA (6-carboxy-tetramethyl-rhodamine).

of each sample [37]. Briefly, the averaged C_T value of the GAPDH transcript was subtracted from the averaged C_T value of the total Arg transcript of a specific cell type in order to obtain the Arg ΔC_T value. The difference ($\Delta\Delta C_T$) between the Arg ΔC_T values in a specific cell type and the Arg ΔC_T value of the LP1 cell line used as a calibrator was determined and expressed as $2^{-\Delta\Delta C_T}$, and represented the fold of Arg expression in relationship to the calibrator. The LP1 cell line was chosen as calibrator from the beginning of the study due to the more mature phenotype among the lymphoid cell lines studied [29]. The relative amount of Arg isoforms was calculated as $2^{-\Delta C_T}$ [37]. A ΔC_T was obtained by subtracting the averaged C_T value of total Arg from that of the target isoform, and was then transformed into $2^{-\Delta C_T}$. This value represents the amount of a single isoform with respect to the total quantity of Arg transcript and is expressed as a percentage. The amplification efficiencies for total Arg, each Arg isoform, and the GAPDH transcripts were determined according to the validation experiments suggested by Applied Biosystems (User Bulletin No. 2) and were approximately equal. In order to confirm the specificity of the PCR reaction, the products of the real-time PCR were electrophoresed on a 1.2% agarose gel.

Western Blotting

The cells were lysed with 1% Triton X-100, 10 mM Tris-HCl PH 7.4, 150 mM NaCl, and the Protease Inhibitor Cocktail (Roche, Mannheim, Germany) as recommended by the manufacturer. The protein concentration was determined by means of a Bio-Rad microassay (Hercules, CA). The lysates (80 μ g) separated in 7.5% polyacrylamide gel electrophoresis were blotted onto nitrocellulose membranes, and stained with Ponceu S in order to show equal lane loading. Western blotting was performed with rabbit polyclonal anti-Arg antibodies [17] directed against the SH2 and SH3 domains (a kind gift of A. Koleske, Yale University, CT). Anti-actin antibodies (Sigma-Aldrich) were used to detect β -actin protein. The detection was performed with secondary antibodies coupled to horseradish peroxidase and a SuperSignal Detection System (Pierce, Rockford, IL).

RESULTS

RT-PCR Qualitative Analysis of Arg Transcripts

The RNA extracted from several cell lines of hematopoietic, epithelial, nervous, and connective origin was analyzed by RT-PCR with two different sets of primers: 41N/42N and 41N/132N. The 41N reverse primer was located on the common exon IV of Arg, the 42N and 132N sense primers were, respectively, located on exons 1A and 1B. Both sets of primers 41N/42N and 41N/132N amplified two

bands of different sizes (respectively 466bp/403bp and 523bp/460bp), demonstrating that the ABL2(ARG) gene is normally expressed in the cells as four different 5'-end transcript isoforms, here called 1A long and short (1AL, 1AS) and 1B long and short (1BL, 1BS) (Figure 2). This was also confirmed with primers spanning different exons and specific for the individual isoform (not shown). These PCR products were all sequenced. The nucleotide sequence showed that 63 bp, which code for 21 amino acids are alternatively juxtaposed to the 1B and 1A first exon. The alternative splicing of the sequence maintain the open reading frame (Figure 2). On the basis of the intron-exon junction of Arg (derived from the NCBI Genome database), the 63 bp sequence was flanked by the consensus sequences of the acceptor (AG) and donor (GT) splice sites. Only the 1AS and 1BL forms were described during the first cloning of Arg cDNA [3] but, in the t(1;12) translocation present in leukemic patients, an identical 63 bp sequence had been found alternatively fused to the Arg common exons in the rearranged transcript [24-26]. However, the lack of 1A and 1B first exons in these rearranged transcripts made it impossible to identify whether this additional splicing event involved both the A and B forms of Arg.

On the basis of the information [17] that the mouse brain Arg sequence specifically excludes an exon encoding amino acids 688 (G) to 791 (S), we used PCR to test the same region of human Arg cDNA. Amplification of the cDNA obtained from the A172 glioblastoma cell line with the 112N/113N primer pair revealed two specific bands of 475 and 166 bp (Figure 2). The presence of different 3'-end forms in cellular RNA samples derived from different cell types was also demonstrated with the 112N/114N and 112N/115N primer pairs that amplified specific bands of 142 and 207 bp, respectively (Figure 2). The isoforms were called C-termini long (CTL) and short (CTS). The sequence revealed that the shorter band was the result of the in-frame loss of a 309 bp fragment encoding amino acids 688 (N) to 790 (G) in the C-termini. The lack of these 103 amino acids affects about half of the F-actin-binding domain [1,15] closest to the Arg kinase domain (Figure 1). The amplification of genomic DNA with the 112N/115N primer pair revealed a single band of 207 bp (as expected from the cDNA sequence), the 112N/114N primers did not reveal any amplified band (not shown). Primer 114N spanned the cDNA sequences that were juxtaposed after the loss of the 309 bp fragment which, in Arg cDNA [3], was delimited by the GGG sequence at both extremities. This GGG sequence also flanks the 5'-end of the 1B exon. On the basis of these data and the NCBI Genome Map Viewer *Homo sapiens* database, Build 34, Version 1, we derived the schema of the intron-exon order and the predicted putative protein isoforms of Arg (Figure 1).

Real-Time PCR Quantitative Analysis of Arg Transcripts

Total Arg transcripts

Real-time PCR with O/P primer pairs and the γ probe complementary to common exons (Table 2) confirmed that the total Arg transcript was more abundant in mature (LP1, Raji, Jurkat) than in immature (ALLP0, Molt-4) cells of the lymphoid leukemic cell lines (Figure 3), as we previously showed with competitive PCR [29]. We also confirmed semiquantitative PCR results showing increased Arg transcript expression in the HL-60 myeloid leukemia cell line which differentiated toward granulocytes or macrophage-like cells [28]. Among the cell lines tested, Arg transcript expression was highest in the A172 glioblastoma cells, in which c-Abl protein was not expressed because of the loss of functionally active germline ABL alleles [38].

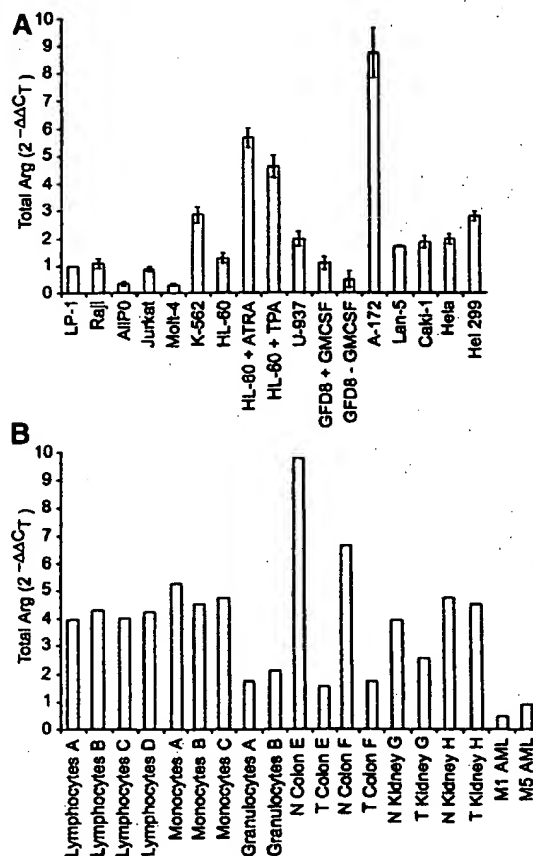


Figure 3. Relative levels of the total Arg transcripts in different human cell types as evaluated by Real-time PCR. The values expressed as $2^{-\Delta\Delta C_T}$ [37] represent the fold of Arg expression in each cell type respect to the LP-1 calibrator cell line, considered as having a value of 1. (A) Cell lines. Mean values of two independent experiments performed in triplicate; the vertical bars indicate the range of variability. (B) Cells and tissues from single individuals denoted by the letters A through H. N, normal; T, tumor; M1, M5, acute myelogenous leukemia FAB types.

The mature lymphocytes, monocytes, and granulocytes from the single normal donors had not only a higher Arg transcript expression than the tumor cell lines of the lymphocytic (LP1, Raji, ALLPO, Molt-4, Jurkat), monocytic (U937), and granulocytic (HL-60, GFD8) lineage, but also than the blasts of myelocytic (M1) and monocytic (M5) spontaneous acute leukemia. A higher expression of Arg had also been observed in normal colon mucosa and in one renal cortex than in the respective carcinomas. On the whole, it seemed that in tumor cells, there was a downexpression of Arg with respect to the normal cells.

5'-End isoforms

The differential quantification of each 5'-end isoform with specific sets of primers and probes (Table 2) demonstrated that the four Arg isoforms were present in different amounts (Figure 4A and B) and had a characteristic expression pattern in the diverse cell types. In hematopoietic cell lines, in Lan-5 neuroblastoma cell line, in lymphocytes, monocytes, and granulocytes, in the M1 and M5 leukemic blasts, the prevailing form was the 1BL followed by its short counterpart (1BS). In these cells, the 1AL form might be quantifiable, while the 1AS form (the only 1A form so far described in literature [3]) was in general the least represented. The most expressed form in A172 glioblastoma and Hel-299 fibroblastic cell lines was 1AL followed by 1AS, although the B forms were quantifiable. In Caki-1 and Hela epithelial cell lines, the 1AL form was the most abundant, but all the other forms (1AS, 1BL and 1BS) were consistently represented accounting for 15%–25% of the total Arg level. A relative distribution of the forms in favor of 1A was found in the two normal renal tissues, this pattern changed with a redistribution of the forms and the prevalence of 1BL in the two renal carcinoma. The forms more represented in the two normal colon mucosa were 1BL and 1BS, this also being true for the tumoral counterpart.

3'-End isoforms

Differential quantification of the 3'-end isoforms showed that all the tested cells contained both CTS and CTL (Figure 4C and D). CTS was prevalent in B lymphoid cell lines, but there was a decrease in CTS with a concurrent increase in CTL in the neoplastic LP1 (plasmacells), Raji (mature B cells), and ALLPO (early pre B cells) cell lines that reflected the difference in maturation. CTS was also greater than CTL in K562 myeloid and Hel-299 fibroblast cells. In the Caki-1 and Hela epithelial cell lines, CTS was more abundant, but the level of CTL was approximately similar. CTL was prevalent in Jurkat and Molt-4 T cell lines, in U937 monocytic cell lines, in A172 glioblastoma, and Lan-5 neuroblastoma cell lines. All donor lymphocytes (mainly T cells),

monocytes, and granulocytes had a prevalence of the CTL form. The two forms were equally expressed in one case of normal colon mucosa, whereas CTL was slightly predominant in the other. In both cases of colon carcinoma, CTS was greater than CTL. The chief form in the two normal renal cortices was represented by CTS. This pattern was unchanged in one of the renal carcinoma, but was inverted in the other.

Real-Time PCR Quantitative Analysis of the Arg Transcript Isoforms in Treated Cells

The HL-60 cells differentiated to granulocytes with 1 μ M ATRA and to macrophage-like cells with 10 nM TPA. The ATRA-treated HL-60 cells stopped growing at d 4 (Figure 5A) showing granulocytic phenotype and morphology [28]. At d 2, the TPA-treated HL-60 cells stopped growing (Figure 5A) and about 60% were adherent to the flask, and were fully viable macrophage-like cells [28]. In the granulocytic differentiation of ATRA-treated HL-60 cells, the expression pattern of the 5'-end isoforms did not change significantly, but a prevalence of the 3'-end CTL form (Figure 5B) as in the granulocytes of volunteer donors (Figure 4B) was observed. In the macrophage-like differentiation of TPA-treated HL-60 cells, the expression profile of the 5'-end isoforms changed dramatically, with a significant increase in the 1A forms and particularly of the 1AL. The 3'-end forms showed redistribution in percentage of CTS and CTL, but the increase in CTL was insufficient to make it more abundant than CTS (Figure 5B).

Given that in HL-60 cells, the differentiation is associated to growth arrest we also investigated the GFD8 cell line in which proliferation blocking could be dissociated from differentiation. The GFD8 cells share properties with early myeloid progenitor cells and are GM-CSF dependent for growth. The presence of GM-CSF does not change the cellular phenotype [34]. Removal of growth factor for 4 d leads to reversible growth arrest (Figure 5A) in the absence of differentiation and without a significant loss of viability [35]. GM-CSF deprivation led to a change in Arg expression at d 4, with the 1A, especially 1AL, forms becoming the most abundant. The 3'-end forms had an increase in the relative difference between CTS and CTL, with CTL remaining preponderant (Figure 5B).

Arg Protein Isoforms Evaluated by Western Blotting

Western blotting analysis of different cell lysates with anti-Arg antibodies revealed a set of bands as previously described [17,29]. The absence or presence of the 21 amino acids of exon II was not sufficient to reveal the different N-terminal isoforms by means of electrophoresis mobility on one-dimension polyacrylamide gel. Thus, the differently sized bands detected by anti-Arg antibodies probably reflected the different sizes of the Arg coding region

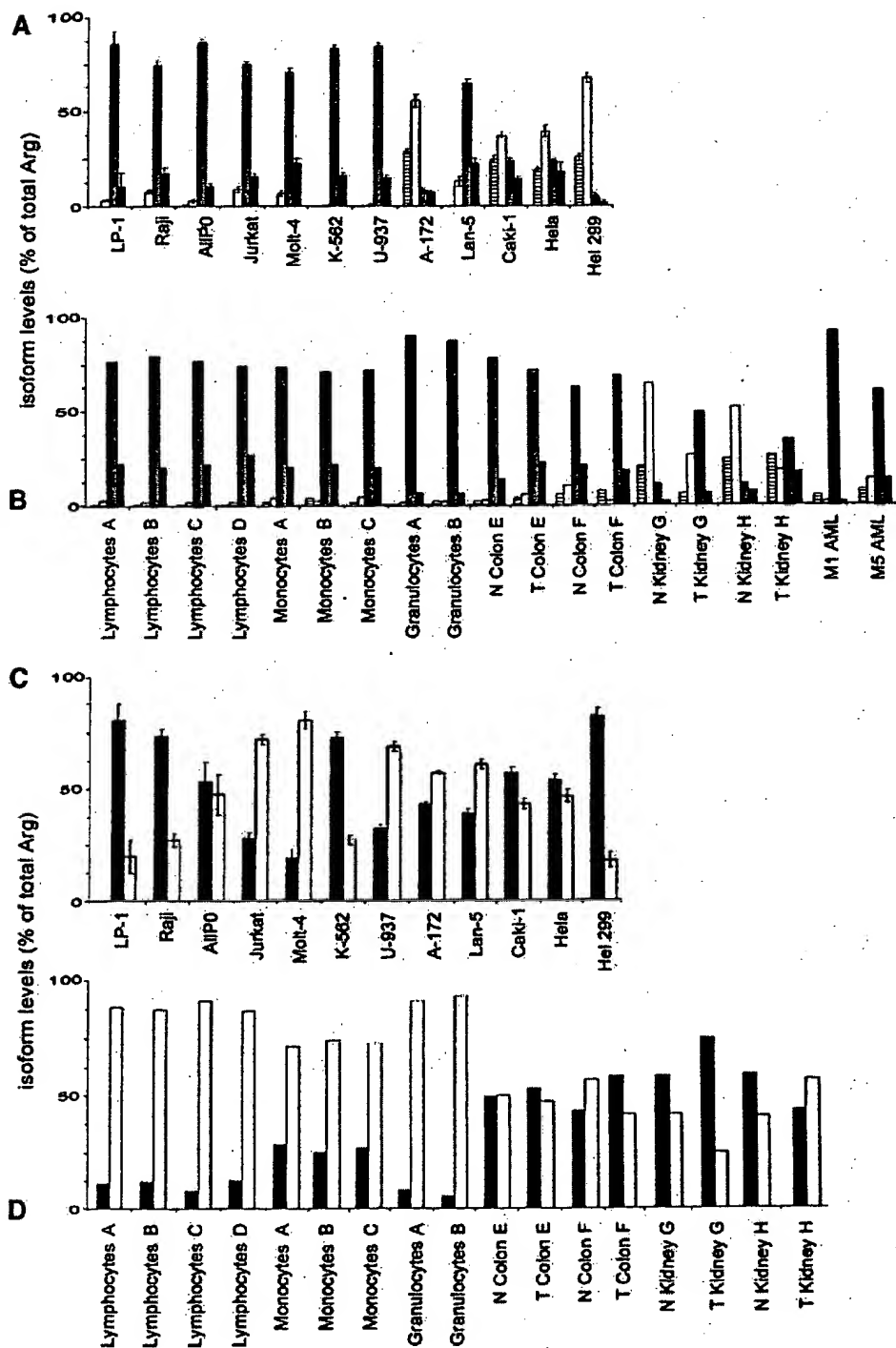


Figure 4. Relative levels of the different isoforms of the Arg transcripts in different human cell types. The values calculated as $2^{-\Delta C_T}$ [37] are reported as the percentage of the individual isoforms with respect to the total Arg transcripts. (A, B) 5'-end isoforms (\square 1AS; \square 1AL; \blacksquare 1BL; \blacksquare 1BS). (C, D) 3'-end isoforms (\blacksquare CTS; \square CTL). (A, C) Cell lines. Mean values of two independent experiments performed in triplicate; the vertical bars indicate the range of variability. (B, D) Cells and tissues from single individuals denoted by the letters A through H. N, normal; T, tumor; M1, M5, acute myelogenous leukemia FAB types.

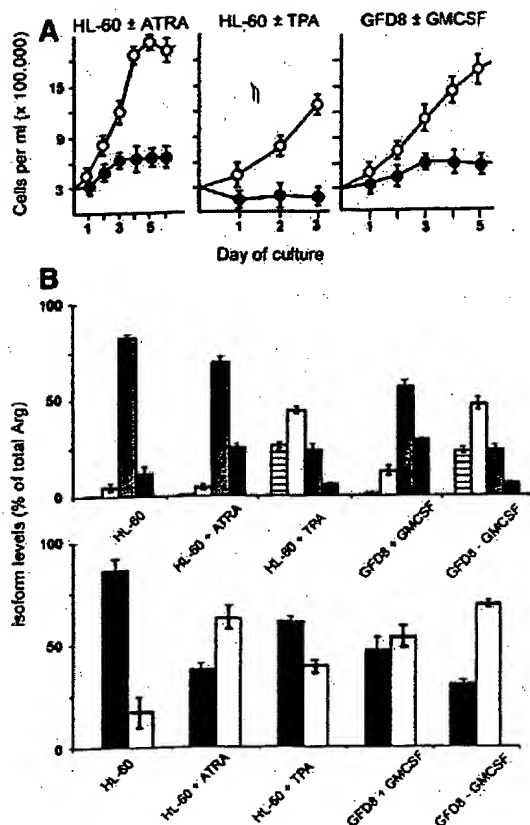


Figure 5. Relative levels of the different isoforms of Arg transcripts in treated cell lines. (A) Growth rate of HL-60 cells, untreated (○) and treated (●) with ATRA or TPA, and GFD8 cells cultivated in the presence (○) or absence (●) of GM-CSF. Exponentially growing cells were plated in 50 ml of medium at a density of 3×10^5 /ml. The total cell number was determined in a Thoma chamber. In the case of the TPA-treated cells, the data refer to the adherent cells detached from the flask after incubation at 37°C with trypsin, and the counted cells are expressed as the number of cells/ml of the initial 50 ml volume. Mean values \pm SD of three independent experiments. (B) HL-60 cells untreated and treated with 1 μ M ATRA for 4 d or 10 nM TPA for 2 d. GFD8 cells cultivated in the presence or absence of 5 ng/ml GM-CSF for 4 d. Top: 5'-end isoforms (\square 1AS; \square 1AL; \blacksquare 1BL; \blacksquare 1BS). Bottom: 3'-end isoforms (\blacksquare CTS; \square CTL). The values calculated as $2^{-\Delta C_T}$ [37] are reported as the percentage of single isoforms with respect to the total Arg transcripts. Mean values of two independent experiments performed in triplicate; the vertical bars indicate the range of variability.

caused by the loss of 103 aminoacids in the C-termini (Figure 6). The protein bands revealed in the different cell types also had a different reciprocal intensity, with a general agreement between band intensity and the reciprocal amount of CTL and CTS transcripts in the different cell lines as quantified by real-time PCR. The shorter protein was greater in K562 cells, and the longer protein in Molt-4 and Lan-5 cells, whereas both proteins were roughly equivalent in the Caki-1 and Hela cells. Instead, in the LP-1 cells the most abundant band was that corresponding to the longer protein, which dis-

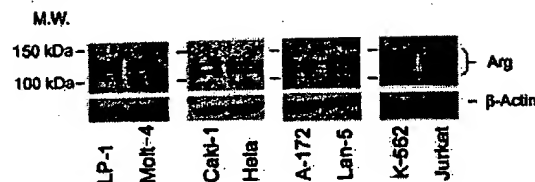


Figure 6. Western blot analysis of 80 μ g of cell lysate from different cell lines separated on 7.5% polyacrylamide gel electrophoresis with polyclonal anti-Arg and anti- β -actin antibodies. The different panels are different blots and the exposure time ranged from 30 to 60 s.

agreed with the transcript data. Finally, it was worth noting that the longer protein seen in A172 cells ran faster than those seen in the Lan-5 and in the other cells. Moreover, the total amount of Arg proteins did not correlate with the Arg mRNA in the A172 cells (Figure 3).

DISCUSSION

During our PCR analysis of the 5'-end of human Arg cDNA, we found two differently sized amplified fragments (long and short) for both 1A and 1B forms. These 1AL, 1AS, 1BL, and 1BS cDNAs diverged in positions that were unique for the long and short forms, thus suggesting that they arose as a result of the combination of two alternative splicing events, one involving the 1A and 1B exons, and the other a second exon. The fact that a 63 bp fragment (which is flanked by a splice acceptor and donor site in the genome) can be alternatively juxtaposed to exons 1A and 1B of Arg suggests that this sequence is the real exon II. In human Arg, the 1A exon therefore ends at amino acid 37 (T) and the 1B exon at amino acid 52 (H), both of which can or cannot be followed by the 21 amino acids of exon II. The long Arg forms (1AL or 1BL) containing exon II are always the most abundant in the cells studied, which suggests that this exon plays an important role in the function of both the A and B forms of Arg proteins, although the exact function of the 21 amino acid sequence has not yet been established. Our findings were also supported by the fact that the translocation involving the ABL2(ARG) gene in leukemia produces the ETV6/ARG fused proteins containing (long form) or lacking (short form) exon II [4-6]. It has been shown that the presence of exon II in the ETV6/ARG long protein leads to more pronounced oncogenic activity [39] in comparison with the short form. The presence of this Arg exon II may therefore be important and also reflect the situation normally existing in the fusion proteins of c-Abl (ETV6/ABL; BCR/ABL), in which the fusion point comes soon after the end of the first exon of c-Abl, but whose second exon does not have any analogy with this second exon of Arg. Furthermore, the N-terminal domains of Arg and c-Abl seem to play an important role in regulating their catalytic activity [40,41], and therefore any study of the kinase activity of Arg

should consider the unique presence of Arg exon II, keeping it distinct from exons 1A and 1B.

Our real-time PCR data showed that the 5'-end short and long forms of Arg mRNA are consistently represented in cells, and four different N-terminal proteins can be predicted from the Arg cDNA sequence even though the difference in N-terminal amino acids is too small to allow their separation by means of one-dimensional gel electrophoresis. The 5'-end isoforms have an expression pattern which differs in the diverse cell types and, in general, it seems that there is a specific prevalence of 1AL and 1AS or 1BL and 1BS isoforms in the various cells: this may suggest, for example, that the 1B first exon has a specific role in the hematopoietic cells in which the 1BL and 1BS isoforms prevail. Moreover, the specific function of the B forms may be differently modulated by the distribution of the long and short forms, which having distinct N-termini may differentially regulate catalytic activity [40,41]. The same can be said for the cells in which the A isoforms predominate.

All of the isoforms may therefore have an as yet unknown functional role, a hypothesis that is also supported by the changes observed during HL-60 cell differentiation and growth arrest of GFD8 cells, and by the fact that in addition to a prevalent form, the others are similarly abundant in some cell types (Caki-1, Hela). The functional impact of the various isoforms can of course be different, depending on their relative abundance. It is worth noting that four mRNA isoforms of mouse c-Abl have been described that diverge at the first exon [12]: the type I and IV isoforms are predominant and translated, and it has been suggested that these two proteins play different roles, with type I being involved in LPS-induced lymphoid differentiation and type IV in apoptosis [42].

The PCR-revealed 309 bp deletion in the last exon of Arg causes the loss of 103 amino acids from the C-termini that affect the actin-binding domain closest to the kinase domain. The two major protein bands revealed by Western blotting with anti-Arg antibodies may represent the two translated isoforms which, on the basis of their amino acid composition, differ by about 10 kDa. The expression pattern of the 3'-end transcript isoforms also varies in the different cell types, and their reciprocal ratio is maintained in the probably translated proteins. The presence of different Arg protein isoforms may have various implications. The C-terminal domains of Arg with diverse actin-binding domains might lead to different interactions with the actin cytoskeletal structure, and to distinct localizations/concentrations of the N-terminal isoforms that might be involved in the activation of different metabolic pathways [40]. A down regulation of Arg expression was observed in the tumors studied as reported in other tumors [7-10]. The Arg isoforms might also play a role in

neoplasms in which an altered Arg expression can be accompanied by variation in the expression pattern of specific Arg isoforms, as we noted in a few tumor cases. Although these data need to be confirmed in a greater number of cases, they open the possibility that any variation of the expression pattern of Arg isoforms might have different and specific effects on morphology and motility or on other specific biological events in tumor cells.

ACKNOWLEDGMENTS

We thank Dr. R. Falbo for revising the manuscript. We thank C. D'Orlando for technical assistance with cell cultures. M.C. and B.E. were supported by M.I.U.R. grants for PhD program. This work was partially supported by A.I.R.C. and M.I.U.R. (ex 60%) grants to R.A.P.

REFERENCES

1. Pendergast AM. The Abl family kinases: Mechanisms of regulation and signaling. *Adv Cancer Res* 2002;85:51-100.
2. Kruh GD, King CR, Kraus MH, et al. A novel human gene closely related to the Abl proto-oncogene. *Science* 1986; 234:1545-1548.
3. Kruh GD, Perego R, Miki T, et al. The complete coding sequence of Arg defines the Abelson subfamily of cytoplasmic tyrosine kinases. *Proc Natl Acad Sci USA* 1990;87: 5802-5806.
4. Cazzaniga G, Tosi S, Aloisi A, et al. The tyrosine kinase Abl-related gene ARG is fused to ETV6 in an AML-M4Eo patient with a t(1;12)(q25;p13): Molecular cloning of both reciprocal transcripts. *Blood* 1999;94:4370-4373.
5. Iijima Y, Ito T, Oikawa T, et al. A new ETV6/TEL partner gene, ARG (ABL-related gene or ABL2) identified in an AML-M3 cell line with a t(1;12)(q25;p13) translocation. *Blood* 2000;95: 2126-2131.
6. Griensinger F, Janke A, Podlenschny M, et al. Identification of an ETV6-ABL2 fusion transcript in combination with an ETV6 point mutation in a T-cell acute lymphoblastic leukaemia cell line. *Br J Haematol* 2002;119:454-458.
7. Chen WS, Kung HJ, Yang WK, et al. Comparative tyrosine kinase profiles in colorectal cancers: Enhanced Arg expression in carcinoma as compared with adenoma and normal mucosa. *Int J Cancer* 1999;83:579-584.
8. Liu LX, Liu ZH, Jiang HC, et al. Profiling of differentially expressed genes in human gastric carcinoma by cDNA expression array. *World J Gastroenterol* 2002;8:580-585.
9. Crnogorac-Jurcevic T, Efthimiou E, Nielsen T, et al. Expression profiling of microdissected pancreatic adenocarcinomas. *Oncogene* 2002;21:4587-4594.
10. Lu TJ, Lu TL, Su JJ, et al. Tyrosine kinase expression profile in bladder cancer. *Anticancer Res* 1997;17:2635-2638.
11. Shtivelman E, Lifshitz B, Gale RP, et al. Alternative splicing of RNAs transcribed from the human abl gene and from the bcr-abl fused gene. *Cell* 1986;47:277-284.
12. Bernards A, Paskind M, Baltimore D. Four murine c-Abl mRNAs arise by usage of two transcriptional promoters and alternative splicing. *Oncogene* 1988;2:297-304.
13. Mysliwiec T, Perego R, Kruh GD. Analysis of chimeric Gag-Arg/Abl molecules indicates a distinct negative regulatory role for the Arg C-terminal domain. *Oncogene* 1996;12: 631-640.
14. Ren R, Ye ZS, Baltimore D. Abl protein-tyrosine kinase selects the Crk adapter as a substrate using SH3-binding sites. *Genes Dev* 1994;8:783-795.
15. Miller AL, Wang Y, Mooseker MS, et al. The Abl-related gene (Arg) requires its F-actin-microtubule cross-linking activity to

- regulate lamellipodial dynamics during fibroblast adhesion. *J Cell Biol* 2004;165:407–419.
16. Wang B, Kruh GD. Subcellular localization of the Arg protein tyrosine kinase. *Oncogene* 1996;13:193–197.
 17. Koleske AJ, Gifford AM, Scott ML, et al. Essential roles for the Abl and Arg tyrosine kinases in neurulation. *Neuron* 1998;21:1259–1272.
 18. Hernandez SE, Settleman J, Koleske AJ. Adhesion-dependent regulation of p190RhoGAP in the developing brain by the Abl-related gene tyrosine kinase. *Curr Biol* 2004;14:691–696.
 19. Moresco EMY, Koleske AJ. Regulation of neuronal morphogenesis and synaptic function by Abl family kinases. *Curr Opin Neurobiol* 2003;13:535–544.
 20. Finn AJ, Feng G, Pendergast AM. Postsynaptic requirement for Abl kinases in assembly of the neuromuscular junction. *Nature Neurosci* 2003;6:717–723.
 21. Burton EA, Plattner R, Pendergast AM. Abl tyrosine kinases are required for infection by *Shigella flexneri*. *EMBO J* 2003;22:5471–5479.
 22. Cao C, Leng Y, Li C, et al. Functional interaction between the c-Abl and Arg protein-tyrosine kinases in the oxidative stress response. *J Biol Chem* 2003;278:12961–12967.
 23. Cao C, Leng Y, Kufen D. Catalase activity is regulated by c-Abl and Arg in the oxidative stress response. *J Biol Chem* 2003;278:29667–29675.
 24. Nishimura N, Furukawa Y, Sutheesophon K, et al. Suppression of Arg kinase activity by STI571 induces cell cycle arrest through up-regulation of CDK inhibitor p18/INK4c. *Oncogene* 2003;22:4074–4082.
 25. Li Y, Shimizu H, Xiang SL, et al. Arg tyrosine kinase is involved in homologous recombinational DNA repair. *Biochem Biophys Res Com* 2002;299:697–702.
 26. Hardin JD, Boast S, Schwartzberg PL, et al. Abnormal peripheral lymphocyte function in c-abl mutant mice. *Cell Immunol* 1996;172:100–107.
 27. Perego R, Ron D, Kruh GD. Arg encodes a widely expressed 145kDa protein-tyrosine kinase. *Oncogene* 1991;6:1899–1902.
 28. Perego RA, Bianchi C, Brando B, et al. Increment of nonreceptor tyrosine kinase Arg RNA as evaluated by semiquantitative RT-PCR in granulocyte and macrophage-like differentiation of HL-60 Cells. *Exp Cell Res* 1998;245:146–154.
 29. Bianchi C, Muradore I, Corizzato M, et al. The expression of the non-receptor tyrosine kinases Arg and C-abl is differently modulated in B lymphoid cells at different stages of differentiation. *FEBS Lett* 2002;527:216–222.
 30. Colotta F, Peri G, Villa A, et al. Rapid killing of actinomycin D-treated tumor cells by human mononuclear cells. *J Immunol* 1984;132:936–944.
 31. Zahler S, Kowalski C, Brosig A, et al. The function of neutrophils isolated by magnetic antibody cell separation technique is not altered in comparison to a density gradient centrifugation method. *J Immunol Methods* 1997;200:173–179.
 32. Bennet JM, Catovsky D, Daniel MT, et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British cooperative group. *Ann Intern Med* 1985;103:620–625.
 33. Gobbi A, Di Bernardino C, Scanziani E, et al. A human acute lymphoblastic leukemia line with the t(4;11) translocation as a model of minimal residual disease in SCID mice. *Leuk Res* 1997;21:1107–1114.
 34. Rambaldi A, Bettoni S, Tosi S, et al. Establishment and characterization of a new granulocyte-macrophage colony-stimulating factor-dependent and interleukin-3-dependent human acute myeloid leukemia cell line (GF-D8). *Blood* 1993;81:1376–1383.
 35. Golay J, Broccoli V, Borleri GM, et al. Redundant functions of B-Myb and c-Myb in differentiating myeloid cells. *Cell Growth Diff* 1997;8:1305–1316.
 36. Kondo M, Kudo K, Kimura H, et al. Real-time quantitative reverse transcription-polymerase chain reaction for the detection of AML1-MTG8 fusion transcripts in t(8;21)-positive acute myelogenous leukemia. *Leuk Res* 2000;24:951–956.
 37. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 2001;25:402–408.
 38. Heinsterkamp N, Morris C, Sender L, et al. Rearrangement of the human ABL oncogene in a glioblastoma. *Cancer Res* 1990;50:3429–3434.
 39. Iijima Y, Okuda K, Tojo A, et al. Transformation of Ba/F3 cells and Rat-1 cells by ETV6/ARG. *Oncogene* 2002;21:4374–4383.
 40. Tanis KQ, Veach D, Duewel HS, et al. Two distinct phosphorylation pathways have additive effects on Abl family kinase activation. *Mol Cell Biol* 2003;23:3884–3896.
 41. Hantschel O, Superti-Furga G. Regulation of the c-Abl and Bcr-Abl tyrosine kinases. *Nat Rev Mol Cell Bio* 2004;5:33–44.
 42. Daniel R, Wong PMC, Chung SW. Isoform-specific functions of c-Abl: Type I is necessary for differentiation, and type IV is inhibitory to apoptosis. *Cell Growth and Differ* 1996;7:1141–1148.



Identification and characterization of multiple isoforms of a murine and human tumor suppressor, *patched*, having distinct first exons[☆]

Kazuaki Nagao^a, Masashi Toyoda^a, Kaori Takeuchi-Inoue^a, Katsunori Fujii^b,
Masao Yamada^a, Toshiyuki Miyashita^{a,*}

^aDepartment of Genetics, National Research Institute for Child Health and Development, 2-10-1 Ohkura, Setagaya-ku, Tokyo 157-8535, Japan

^bDepartment of Pediatrics, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan

Received 3 September 2004; accepted 23 November 2004

Available online 11 January 2005

Abstract

Mutations in mouse and human *patched* (*PTCH*) genes are associated with birth defects and cancer. *PTCH*, a 12-pass transmembrane protein, is a receptor for Sonic hedgehog (Shh) signaling proteins. Shh proteins activate transcription of target genes, including *PTCH*, via GLI transcription factors. Here we identified seven and five isoforms of human and mouse *PTCH* mRNA, respectively, which are generated by the complex alternative use of five exons as the first exon (exons 1a to 1e in the 5'-to-3' order). Although expression profiles of these isoforms were highly variable among human tissues, three of them, *PTCHa*, *PTCHb*, and *PTCHd*, were predominantly expressed in most tissues, *PTCHd* being most ubiquitous. In contrast, *PTCHb* was always predominant and reached a maximum at E10.5 during mouse development. These three mRNA isoforms encode three *PTCH* proteins with distinct N-termini, *PTCH_L*, *PTCH_M*, and *PTCH_S*. The expression of these three isoforms was regulated by GLI transcription factors, and at least two functional GLI-binding sequences were identified, one in exon 1a and the other between exon 1a and exon 1b. *PTCH_L* and *PTCH_M* were equally active in terms of suppressing GLI-mediated transcription and inducing apoptosis. *PTCH_S* protein (encoded by *PTCHd*), lacking the first transmembrane domain, was more unstable than the other two, resulting in a reduced activity. This study may shed light on the mechanism whereby a single *PTCH* gene plays a role in both tumor cell growth and embryonic development.

© 2004 Elsevier Inc. All rights reserved.

Keywords: *Patched*; Sonic hedgehog; Basal cell carcinoma; Medulloblastoma; Alternative splicing

The Sonic hedgehog (Shh) signaling cascade is pivotal to embryonic development, because holoprosencephaly (HPE), characterized by a failure of the forebrain to separate completely into hemispheres, and HPE-like abnormalities are associated with a loss of Shh function in humans and in mice [1–3]. The role of the Shh pathway in tumorigenesis was also established with the discovery that inactivating mutations in the *Patched* (*PTCH*) gene, which encodes one component of the Shh receptor, are responsible for the inherited cancer predisposition disorder known as Gorlin's

or nevoid basal cell carcinoma syndrome (NBCCS) [4,5], as well as sporadic basal cell carcinomas (BCCs) and medulloblastomas [6–8]. NBCCS is an autosomal dominant neurocutaneous disorder characterized by developmental abnormalities such as palmar and plantar pits, jaw cysts, calcification of the falx cerebri, and skeletal anomalies and also by a predisposition to cancers such as BCC and medulloblastoma [9]. Familial and sporadic BCCs display loss of heterozygosity in this region, consistent with *PTCH* being a tumor suppressor gene [6,10]. In addition, activating mutations in *Smoothened* (*Smo*), also encoding another component of the Shh receptor, have been detected in BCCs [11], further emphasizing the importance of this pathway in tumor development. More importantly, the recent finding that this pathway is essential for growth of a wide range of tumor types not associated with NBCCS, such as lung

[☆] Sequence data from this article have been deposited with the GenBank Library under Accession Nos. AB164615, AB164616, and AB189436–AB189442.

* Corresponding author. Fax: +81 3 5494 7035.

E-mail address: tmiyashita@nch.go.jp (T. Miyashita).

cancers or digestive tract tumors, sheds light on potential new diagnostic and therapeutic approaches [12–14].

PTCH, a 12-pass transmembrane protein, is the ligand-binding component of the Shh receptor complex. In the absence of Shh binding, PTCH is thought to hold Smo, a 7-pass transmembrane protein, in an inactive state and thus inhibit signaling to downstream genes. Upon the binding of Shh, the inhibition of Smo is released and signaling is transduced, leading to the activation of target genes by the Gli family of transcription factors [15]. The transcription of *PTCH* itself is induced by Shh pathway activity [16], thus generating a negative feedback loop, which may play an important role in tumor suppression by inhibiting a sustained activation of the pathway.

Hahn et al. predicted that there are three different forms of the PTCH protein present in humans: the ancestral form and two human-specific forms [4]. Recently, a detailed characterization of three alternative first exons was reported [17]. However, our study using the 5' rapid amplification of cDNA ends (5'RACE) technique revealed the existence of an additional first exon and unexpectedly complex splicing between the first and the second exons that is evolutionarily conserved across species. Therefore, the characterization of several potential forms of the PTCH protein may reveal the mechanism whereby a single *PTCH* gene could play a role in different pathways, and the determination of the regulation of different splice forms of *PTCH* mRNA may shed light on the apparent role of the gene in tumor cell growth as well as embryonic development. Here we

characterize multiple isoforms of *PTCH* in humans and mice and discuss the functions of their products, expression profiles, and transcriptional regulation.

Results

Isolation of isoforms of human and mouse *PTCH*

PTCH is a multiexon gene comprising 23 exons distributed over a region of ~70 kb. To date, three cDNA sequences encoding the human *PTCH* gene's first exon have been reported and named exons 1, 1A, and 1B [17], and another exon has recently been deposited with GenBank (exon 1a described below, GenBank Accession No. BC043542). In contrast, only a single mRNA species of *PTCH* has been reported in mice [18] (GenBank Accession No. U46155). Due to the use of alternative exons, several mRNA isoforms are generated. On the basis of this background we performed a comprehensive analysis of the 5' structure of mRNA species derived from the human *PTCH* gene employing the 5'RACE technique. Sequencing of 31 RACE clones revealed an additional alternative first exon (exon 1c described below, submitted to GenBank as Accession No. AB189438) and complex splicing between the first and the second exon. Using a genomic sequence containing the *PTCH* gene (GenBank Accession No. AL161729), the precise genomic organization of the human *PTCH* gene was determined as shown in Fig. 1. For the sake

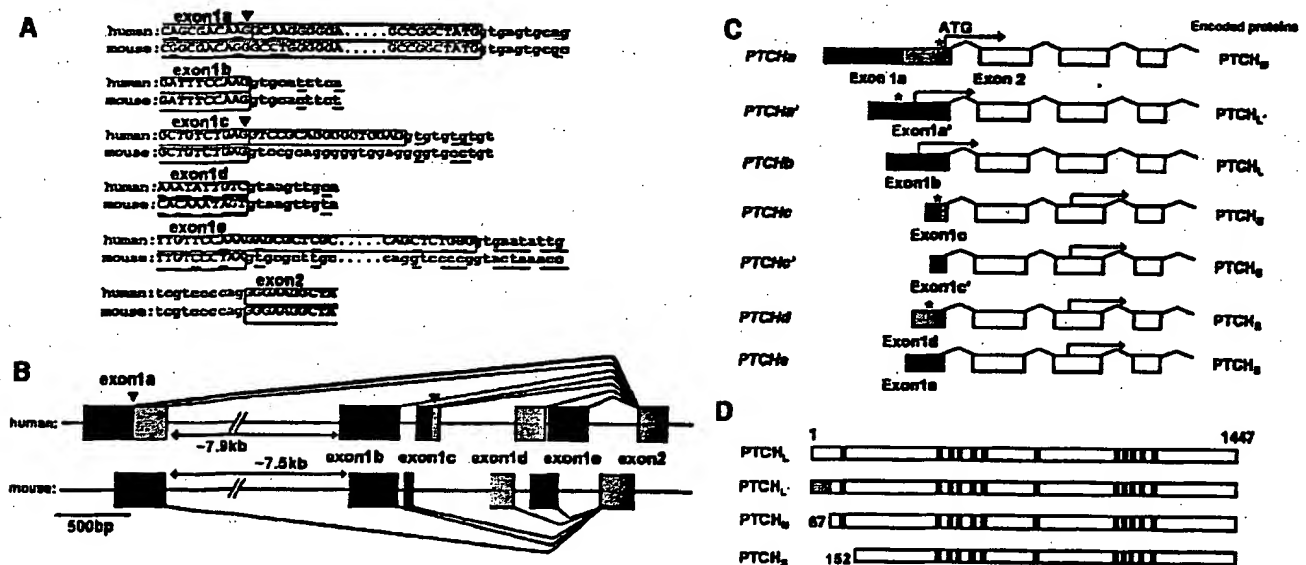


Fig. 1. Identification of human and mouse *PTCH* isoforms. (A) Comparison of human and mouse exon-intron boundaries. Upper- and lowercase letters indicate exon and intron sequences, respectively. Nucleotides not conserved between the two species are underlined. Alternative splice donor sites are indicated by arrowheads. (B) 5' region of human and mouse *PTCH* gene structure. The 5' ends of the mouse first exons have not been determined. (C) 5' structure of *PTCH* isoforms. The positions of the first methionine codons and in-frame stop codons are indicated by arrows and asterisks, respectively. In four of seven mRNAs, in-frame stop codons were identified. The first in-frame methionine codon could be determined in the other three transcripts since the 5'RACE system we employed amplifies only full-length transcripts [47]. (D) *PTCH* protein isoforms encoded by mRNA species described in (C). Numbers refer to amino acid positions relative to the first methionine of PTCH_L. The positions of the 12 transmembrane regions are indicated by filled boxes. PTCH_L has 65 unique amino acid residues at the N-terminus depicted with a shaded box.

of simplicity, we named the first exons exon 1a to 1e on the basis of their 5'-to-3' order. Thus, exons 1b, 1d, and 1e are the former exons 1B, 1, and 1A, respectively. In addition to multiple first exons, we found that alternative 5' splice sites allow the shortening of exons 1a and 1c, generating exons 1a' and 1c' (Fig. 1C). The complex alternative splicing described above thus generates up to seven mRNA species, each with its own distinct 5' sequence (Figs. 1B and 1C). RT-PCR using isoform-specific forward primers for each alternative exon 1 and a common reverse primer for exon 2 indeed validated the existence of the seven different mRNAs. These mRNA isoforms encode four PTCH proteins termed PTCH_L, PTCH_{L'}, PTCH_M, and PTCH_S (Figs. 1C and 1D). PTCH_S is an N-terminally truncated PTCH protein that lacks the first transmembrane domain (Fig. 1D). Although only a single species of *PTCH* mRNA has been reported in mice, a comparison of the human *PTCH* genomic sequence with the mouse sequence (NCBI Locus NT_039587) suggested the existence of multiple first exons. In this study, mouse and human *Patched* genes are collectively referred to by the human nomenclature (*PTCH*, whereas mouse *Patched* is often called *Ptc*). RT-PCR using the forward primers constructed at mouse putative first exons and reverse primers at exon 2 demonstrated that most of the *PTCH* isoforms found in humans are indeed conserved in mice. At least in mouse P19 cells and several mouse tissues from which total RNA was extracted, *PTCHa'* and *PTCHc* have not been identified and the splice donor site at exon 1e was different from that of humans (Fig. 1A). All exons were flanked by splice junctions that conformed to the consensus GT-rule except for exon 1a'-exon 2 in humans, in which the GC-AG intron was observed. GC-AG introns are occasionally found and processed by the same splicing pathway as conventional GT-AG introns [19].

Expression profiles of three isoforms of *PTCH* in various tissues

Selective usage of the 5'-most exons suggests a complex tissue-specific transcriptional regulation. Therefore, to investigate the expression profiles of *PTCH* isoforms, RT-PCR was performed with isoform-specific primers for the first alternative exons using total RNA from a panel of human tissues, and profiles were analyzed with an Agilent 2100 bioanalyzer. As shown in Fig. 2A, *PTCH* was expressed in a wide range of human tissues. However, the levels of total *PTCH* RNA varied among human tissues. For example, the heart and liver showed low levels of expression, which is largely consistent with previous reports on human and mouse *PTCH* expression [18,20]. Expression profiles of the *PTCH* isoforms were also highly variable among tissues. While *PTCHd* (encoding PTCH_S) was widely expressed, the expression of *PTCHa* (encoding PTCH_M) and *PTCHb* (encoding PTCH_L) was relatively restricted. For example, *PTCHb* was expressed in all the

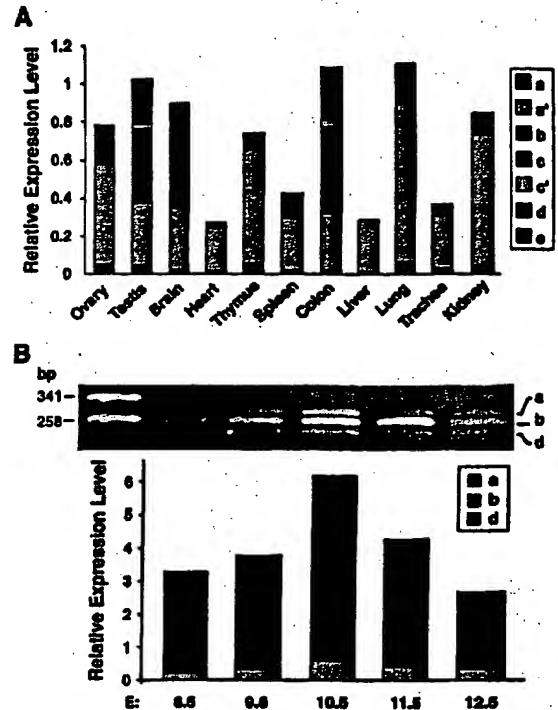


Fig. 2. Expression profiling of *PTCH* isoforms. (A) RT-PCR analysis of expression profiles in various tissues. Total RNA obtained from a panel of human tissues was subjected to RT-PCR. Forward primers specific to each of the first exons and a common reverse primer for exon 2 were synthesized and used for PCR. The RT-PCR products were quantified with an Agilent 2100 bioanalyzer. *PTCH* expression levels were normalized to those of *GAPDH*. Exons with relative expression levels lower than 0.007 do not appear in the graph. (B) RT-PCR analysis of expression profiles in various mouse developmental stages. Total RNA obtained from mouse embryos at various developmental stages was subjected to RT-PCR using mouse-specific primers. Mean *PTCH* expression levels normalized to β -actin expression are presented at the bottom ($n = 2-4$).

analyzed tissues apart from liver, while the expression of *PTCHa* was more restricted, showing virtually no expression in the heart, thymus, liver, and trachea. The other *PTCH* isoforms using exons 1a', 1c, 1c', and 1e were found to be expressed at very low levels if at all throughout the tissues. Therefore, we focused on *PTCHa*, *PTCHb*, and *PTCHd* in further experiments. Since Shh signaling plays a key role in embryonic development, we next investigated the expression profile in mouse embryogenesis. Consistent with a previous report, the expression of *PTCH* reached a maximum at E10.5, at which point the limb buds become increasingly prominent, and declined thereafter [18]. Notably, in contrast to human adult tissues, the expression of *PTCHb* was always prevalent during embryonic development (Fig. 2B).

Transcriptional regulation of *PTCH* isoforms by *GLI*

It is well known that *PTCH* itself is one of the target genes in the Shh signaling network creating a negative feedback loop and a balance via the antagonism of Shh and

PTCH. Even though the GLI proteins may well not be the only mediators of Shh signaling, the overwhelming majority of available data on insects and vertebrates indicates a central role for GLI proteins in regulating the mediation and interpretation of Shh signals. As shown in Fig. 3, the expression of all three *PTCH* isoforms was elevated by GLI1 in the cell lines we employed. However, a closer observation revealed slight differences in the degree of induction. For example, *PTCHd* and *PTCHb* were more strongly upregulated by GLI1 in 293T and HSC-2 cells, whereas the induction of *PTCHa* was more evident than that of *PTCHb* or *PTCHd* in Ho-1-u-1 and LK-2 cells, indicating cell type-specific regulation of the isoforms.

PTCH promoter has functional GLI-binding sites

The *Drosophila patched* gene (*ptc*) has a cluster of three GLI consensus binding sites (5'-TGGGTGGTC-3' or 5'-GACCA-3') [21] in the promoter region that is required for the reporter gene expression in response to Hedgehog (Hh) activity [22]. Recently, it was reported that the transcriptional regulation of *PTCH* by Shh signaling was mediated by a single GLI-binding site located ~400 bp upstream of exon 1b (GLI-BS1 in Fig. 4A) [23]. However, sequencing farther upstream indicated the presence of even two more consensus GLI-binding sequences not reported previously (GLI-BS2 and GLI-BS3 in Fig. 4A, -3965 and -8283 bp relative to the reported transcription start site of exon 1b, respectively). The mouse upstream sequence also contained three putative consensus GLI-binding sites and

the sequences around these sites were strikingly conserved (Fig. 4B). This suggests that two upstream consensus GLI-binding sequences, as well as a reported one, act as GLI-responsive elements. To test this assumption, genome fragments containing GLI-BS1, GLI-BS2, and GLI-BS3 were inserted into a luciferase construct (pGV-PTCH1, pGV-PTCH2, and pGV-PTCH3, respectively). Cotransfection of the GLI1 expression plasmid with pGV-PTCH1 enhanced the luciferase activity in SH-SY5Y cells (Fig. 4C), confirming a previous report. In addition, as anticipated, GLI1 expression also enhanced the luciferase activity when cotransfected with reporter constructs containing upstream GLI-binding sequences (pGV-PTCH2 and pGV-PTCH3). To confirm that these sites are really responsible for the GLI-mediated activation, a mutation with four nucleotide substitutions was introduced into GLI-binding sequences (5'-TAGTGGATC-3' or 5'-GATCCACTA-3', mutated nucleotides in italic), generating the constructs pGV-PTCH1mt, pGV-PTCH2mt, and pGV-PTCH3mt. The introduction of these mutations into the putative GLI-binding sites indeed abolished the elevation of luciferase activity induced by GLI1. Furthermore, the 1.1-kb mouse fragment containing GLI-BS1 showed a similar response to GLI1 expression (pGV-mPTCH) (Fig. 4C), suggesting that the mechanism by which *PTCH* expression is regulated by the Shh signaling pathway is conserved.

We also examined whether GLI protein could physically associate with putative GLI-binding elements in *PTCH* in vitro and in vivo. First, we tested these sites in an electrophoretic mobility shift assay. As shown in Fig. 4D, when GST-GLI3 fusion protein was incubated with a wild-type DNA probe containing a putative GLI consensus sequence in the promoter region, a complex with a shift in gel mobility was detected (lane 3). In contrast, substitution of GST nonfusion for GST-GLI3, or mutant DNA probe with the same nucleotide substitutions as described above for the wild-type sequence, resulted in a failure to detect a complex whose mobility was altered in these assays (lanes 2 and 6). Moreover, the DNA-protein complex was abolished by competition with an unlabeled oligonucleotide containing the GLI site, but not by a mutated oligonucleotide, demonstrating the specificity of the complex formation (lanes 4 and 5). GST-GLI3 also bound specifically to two more upstream sequences with a GLI-binding consensus sequence (lanes 9 and 15) in vitro.

To determine whether the GLI protein occupies these sites in vivo, we used a chromatin immunoprecipitation (ChIP) assay to analyze lysates extracted from 293T cells transfected with a plasmid to express Flag-GLI1. The genomic fragments including GLI-BS1 and GLI-BS3 were specifically precipitated as a GLI-DNA complex with an anti-Flag antibody (Fig. 4E, lanes 3 and 11), while GLI-BS2 was barely coimmunoprecipitated (lane 7). As controls, the same fragments were not precipitated when cells were transfected with a construct for Flag tag or the lysates were incubated with an anti-Myc antibody (lanes 2, 4, 10, and

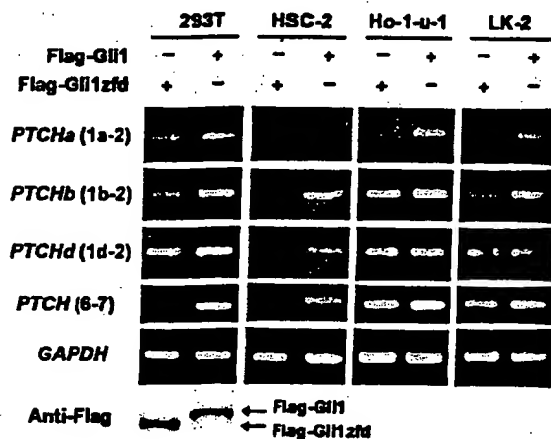


Fig. 3. Transcriptional regulation of three *PTCH* isoforms. Cell lines indicated at the top were transfected with the expression plasmid pSRα-Flag-Gli1 or pSRα-Flag-Gli1zfd. pSRα-Flag-Gli1zfd is a plasmid for a mutant GLI1 lacking the zinc finger domain [43] used as a negative control for pSRα-Flag-Gli1. Cells were cultured in 0.5% FCS for 16 h after the transfection and total RNA was extracted from the transfected cells and subjected to RT-PCR. Forward and reverse primers were constructed for the exons indicated in parentheses. *PTCH* (6–7) indicates the overall *PTCH* expression because exons 6 and 7 are used regardless of the isoform. The expression of Flag-tagged GLI1 proteins was confirmed by immunoblotting using anti-Flag antibody (Anti-Flag).

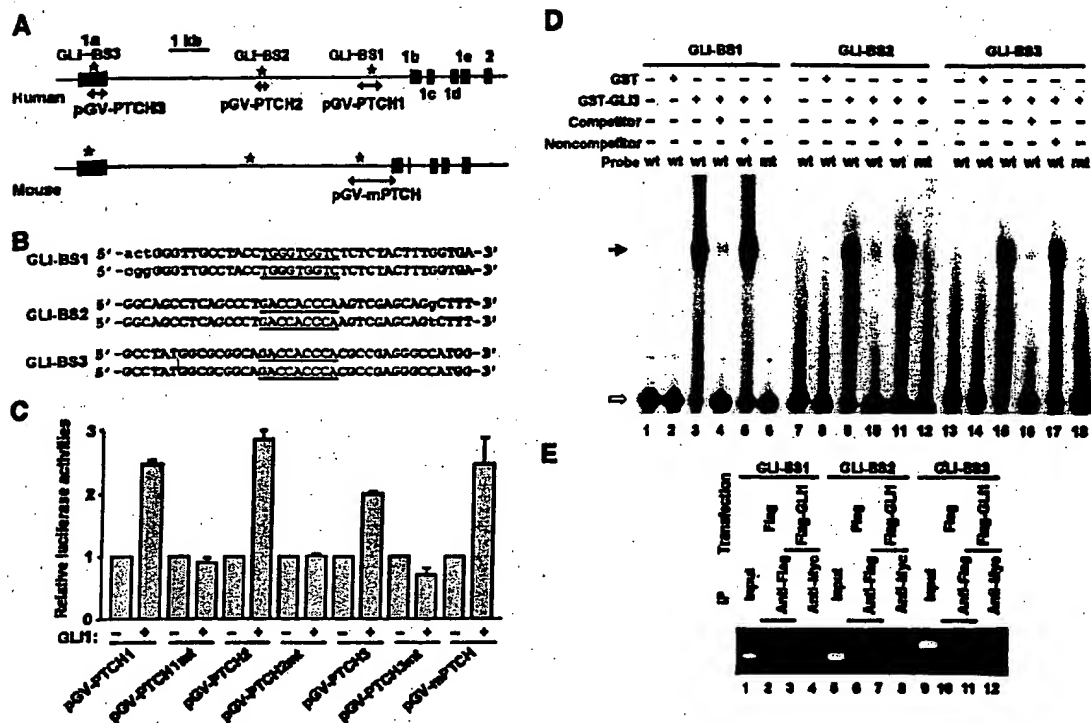


Fig. 4. Transcriptional regulation of *PTCH* isoforms. (A) Comparison of human and mouse genomic structures. Black boxes indicate locations and relative sizes of exons. Asterisks indicate the positions of three putative GLI-binding sequences (5'-TGGGTGGTC-3' or 5'-GACCACCCA-3'). DNA fragments inserted into luciferase vectors to make the reporter gene constructs are indicated by arrows. The names of the resulting constructs are indicated below. (B) Comparison of human (top) and mouse (bottom) GLI-binding sequences. Consensus GLI-binding sequences are underlined. Lowercase letters indicate nucleotides not conserved between the two species. (C) The *PTCH* promoter is GLI responsive. SH-SY5Y cells were cotransfected with various reporter gene constructs as indicated with or without pSRα-Flag-GLI1. Cells were cultured in 0.5% FCS for 16 h after the transfection and then harvested for the luciferase assay. Firefly luciferase activity was normalized by *Renilla* luciferase activity from a cotransfected pRL-SV40 and is indicated relative to the activity of the same reporter without pSRα-Flag-GLI1. The total amount of transfected DNA was adjusted using pcDNA3.0. Data are representative of three experiments with similar results. (D) GLI protein can bind in vitro to an oligonucleotide probe representing the *PTCH* gene region. Recombinant GST or GST-GLI3 protein was incubated with ³²P-labeled oligonucleotide DNA probes containing a putative GLI-consensus sequence (wt) or a mutated version with four nucleotide substitutions (mt), together with or without a 50-fold molar excess of cold competitor containing the GLI site (competitor) or its mutant (noncompetitor). DNA-protein complexes were size fractionated in a nondenaturing polyacrylamide gel and were detected by autoradiography. The positions of the free probe and the shifted complexes are indicated by the open and closed arrows, respectively. (E) Identification of GLI-binding region in vivo. ChIP assay was performed with genomic fragments including the putative GLI-binding consensus sequence indicated at the top. Chromatin from 293T cells transfected with pCI-Flag (lanes 2, 6, 10) or pFlag-GLI1 (lanes 3, 4, 7, 8, 11, 12) was immunoprecipitated with anti-Flag antibody (lanes 2, 3, 6, 7, 10, 11). PCR amplification was performed with corresponding templates. Input represents a portion of the sonicated chromatin before immunoprecipitation. Anti-Myc antibody was used as a negative control (lanes 4, 8, 12).

12). Taken together, our data show that at least GLI-BS1 and GLI-BS3 are involved in GLI-mediated *PTCH* expression. In contrast, GLI-BS2 is not accessible to GLI in vivo, probably due to a higher genomic structure, although the accessibility may be cell-type dependent.

Functional analysis of three isoforms of *PTCH*

In 293T cells, overexpression of *PTCH* protein causes apoptosis and inhibition of cell proliferation [24,25]. Thus, it is expected that there is a basal level of leakage activity of Smo that excess *PTCH* prevents in the apparent absence of Shh. The fact that cyclopamine has a proapoptotic effect in these cells supports this possibility (discussed below). On the basis of this background, we performed a functional analysis of the *PTCH* isoforms using a GLI-responsive luciferase reporter in 293T cells. Luciferase activities were

suppressed when 293T cells were transfected with plasmids for *PTCH_L* and *PTCH_M* but not with an empty vector, pcDNA3.0 (Fig. 5A). This suppression was not observed when cells were transfected with the plasmid for *PTCH_{ΔC}* which encodes only 194 N-terminal amino acid residues, indicating the specificity of the results. To investigate the function of *PTCH* in vivo, *PTCH* was transiently expressed in 293T cells. As expected, *PTCH_L* and *PTCH_M* induced apoptosis in 293T cells as measured by assessing the sub-G0/G1 population (Fig. 5B). However, they were not as potent as cyclopamine, a well-known inhibitor of Shh signaling [26]. This is probably, at least in part, due to the presence of untransfected cells. Interestingly, in contrast to *PTCH_L* and *PTCH_M*, *PTCH_S* did not significantly suppress GLI-responsive luciferase activity or induce apoptosis, implying that this isoform does not have the expected function of a *PTCH* protein or the expression level of this isoform

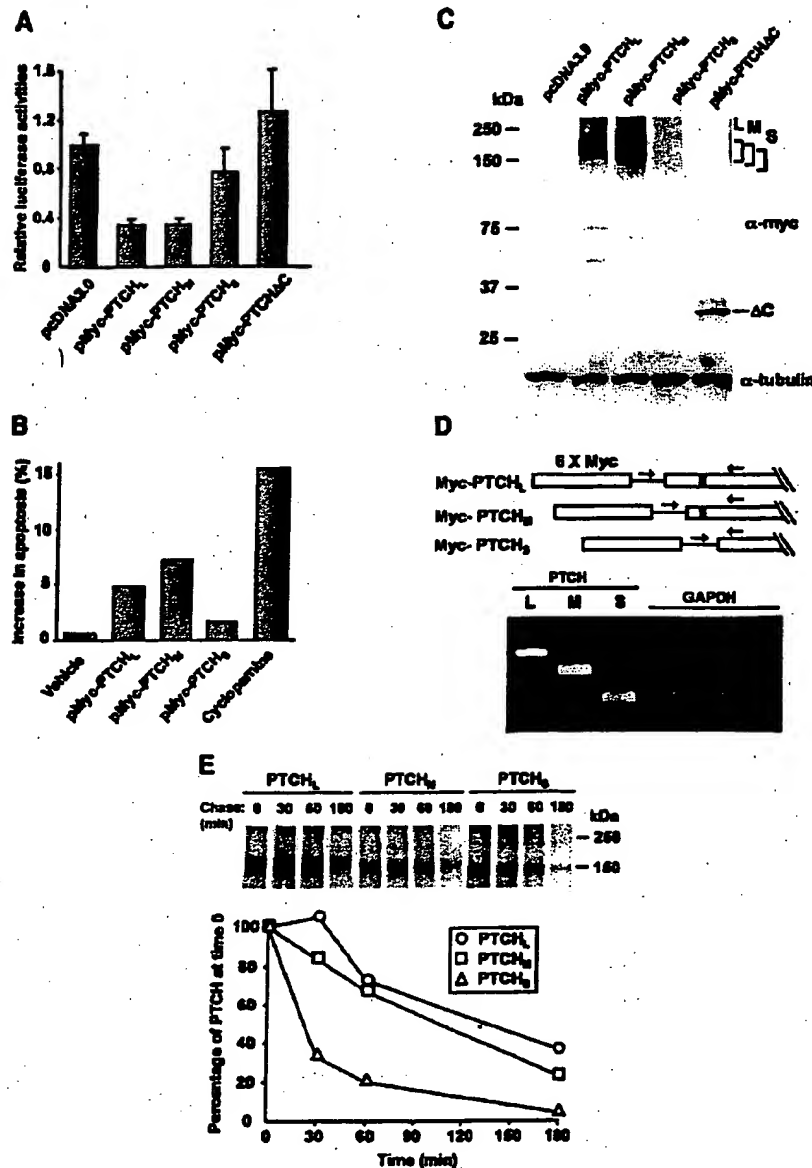


Fig. 5. Functional analysis of PTCH isoforms. (A) Inhibition of GLI-responsive luciferase activity by PTCH. 293T cells were transfected with various expression plasmids as indicated together with 8 × GLI-Luc containing eight GLI-binding sites and LTR-LacZ. After the transfection, cells were cultured in 0.5% FCS for 16 h and then harvested for the luciferase assay. Firefly luciferase activity was normalized to β-galactosidase activity from a cotransfected LTR-lacZ vector. Data are representative of three experiments with similar results. (B) PTCH-induced cell death as measured based on DNA content. 293T cells were transfected with plasmids for PTCH or treated with cyclopamine or vehicle alone (ethanol). The induction of apoptosis was assessed by the increase in the subG0/G1 population compared with mock-transfected cells. (C) Protein levels of expressed genes. Cell lysates were obtained from 293T cells transfected with indicated plasmids and subjected to immunoblotting with an anti-c-Myc antibody. Tubulin is a loading control. The molecular weights of the four PTCH protein products predicted from the composition of amino acid residues, including the Myc tag, are as follows: PTCH_L, 172 kDa; PTCH_M, 163 kDa; PTCH_S, 154 kDa; PTCHΔC, 32.2 kDa. (D) RT-PCR analysis of expressed genes. Total RNA was extracted from 293T cells transfected with plasmids for each isoform of PTCH and RT-PCR analysis was performed using primers depicted at the top. A forward primer was constructed in the linker region between the Myc tag and PTCH and a reverse primer in exon 2. Filled boxes indicate the position of the first transmembrane domain. GAPDH is an internal control for RT-PCR. (E) Metabolic labeling of the PTCH proteins. 293T cells transfected with a construct for PTCH were pulse-labeled with [³⁵S]methionine and chased for the indicated periods. ³⁵S-labeled PTCH was immunoprecipitated, detected by autoradiography (top), and then quantified by phosphorimaging. Levels of labeled PTCH are plotted relative to the amount present at time 0 (bottom).

is too low to cause these changes. To examine these possibilities, we first investigated the protein levels of each PTCH isoform by immunoblotting. Compared with PTCH_L, PTCH_M, and PTCHΔC, the protein level of PTCH_S was markedly reduced (Fig. 5C). The diffuse migration of PTCH

proteins is thought to be due to glycosylation as reported [27,28]. However, when RT-PCR was performed to analyze mRNA levels, these three isoforms were found to be expressed at comparable levels (Fig. 5D). These results indicate that the stability of PTCH_S protein is compromised.

To determine whether the reduced activity of $PTCH_S$ was due to decreased protein stability, we measured the half-life of the three isoforms. 293T cells transfected with a plasmid for each isoform were metabolically labeled with [35 S]-methionine and then incubated with excess unlabeled amino acids for various lengths of time. $PTCH$ proteins were immunoprecipitated and size-separated by SDS-PAGE. As shown in Fig. 5E, Myc-tagged $PTCH$ proteins were visualized at a point corresponding to approximately the same size as that detected by immunoblotting. Following a 180-min chase, 36 and 23% of de novo synthesized $PTCH_L$ and $PTCH_M$, respectively, remained in 293T cells. Half-lives were calculated as 115 and 83 min, respectively. In contrast, the degradation of $PTCH_S$ was considerably accelerated, such that 5% of the protein remained at 180 min (half-life 26 min). These results indicated that $PTCH_S$ is an unstable protein compared with $PTCH_L$ and $PTCH_M$.

Discussion

Alternative pre-mRNA splicing is an important mechanism for generating protein diversity and may explain in part how mammalian complexity arises from a surprisingly small complement of genes. It also plays important roles in development and disease. A recent study estimated that greater than 55% of human genes are alternatively spliced [29] and that about 10% of the mutations in the human genome affect the canonical splice site sequence [30]. In particular, isoforms of genes with alternative first exons may have distinct mechanisms of expression. For example, the *DSCR1* (Down syndrome candidate region 1)/*MCIP1* (modulatory calcineurin-interacting protein 1) and *nNOS* (neuronal nitric oxide synthase) genes have four and eight alternative first exons, respectively, and are subjected to a distinct expressional regulation by separate promoters [31,32].

In this study, we identified and characterized five alternative first exons in both human and mouse *PTCH* genes encoding four protein species. Thus, arguably, *PTCH* is one of the most complex human genes in terms of diversity at the 5' end. The transcription of all major isoforms was upregulated by GLI1, an upstream transcription factor in the Shh pathway, although the degree of activation was cell type-specific. Unlike *Drosophila ptc* in which only a single transcript has been reported and whose promoter has a cluster of three GLI-binding consensus sequences in a 130-bp region [22], human and mouse *PTCH* have three consensus sequences dispersed over 7.5 kb between exon 1a and exon 1b (Fig. 4A). Since exons 1b, 1c, 1d, and 1e are located close to each other, it is likely that *PTCH* isoforms except *PTCHa* are regulated by at least partially overlapping promoters, including GLI-BS1 in Fig. 4A. In contrast, exon 1a is located ~8 kb upstream of exon 1b and one of the GLI-binding sites is located inside exon 1a and the other two are located far downstream. No

GLI-binding consensus sequence was found in the promoter region of *PTCHa* (i.e., upstream of exon 1a), at least not up to the 40 kb position. Thus, taking our results with the ChIP assay into consideration, it is likely that the two GLI-binding sequences, one in exon 1a and the other far downstream of exon 1a (GLI-BS3 and GLI-BS1 in Fig. 4A, respectively), are responsible for the GLI-mediated regulation of *PTCHa*. This is not unexpected because *hepatocyte nuclear factor-3 β* , another target gene of Shh signaling, has a GLI-binding site 3' of the transcription unit and this site is essential for the response to Shh [33]. Although NBCCS families who show linkage to chromosomal regions other than 9q22.3–q31, to where *PTCH* has been mapped, have not been reported, a considerable number of NBCCS patients do not have mutations within the coding region of *PTCH* [34–36]. Therefore, taking our results into account, it is warranted to examine mutations in GLI-binding sequences using samples from such patients. Interestingly, *PTCH2*, another homologue of the *Drosophila Hh* gene, whose mutations are found in BCC and medulloblastoma [37], also has a GLI-binding consensus sequence ~470 bp upstream of the first methionine codon (based on the genomic sequence, AL136380), indicating that *PTCH2* is another target gene of the Shh pathway. Supporting this notion, *PTCH2* is upregulated in basal cell carcinoma in which Shh signaling is activated [38].

$PTCH_L$ and $PTCH_M$ were equally potent in terms of suppressing GLI-mediated transcription or inducing apoptosis. In contrast, the $PTCH_S$ protein was less potent due to its instability. Amino acid residues 101–119 of $PTCH_L$ and 35–53 of $PTCH_M$ comprise the first transmembrane domain, which is absent in $PTCH_S$ because it starts with Met¹⁵² in $PTCH_L$ (Fig. 1D). This probably explains why $PTCH_S$ is unstable. However, *PTCH_S* was more ubiquitously expressed throughout adult tissues than the other two, implying that, despite its instability, $PTCH_S$ may be important for tissue homeostasis or tumor suppression. It is possible that a certain extracellular stress or stimulus such as the binding of Shh may stabilize $PTCH_S$. In contrast, the expression of $PTCH_L$ was always predominant during embryonic development, indicating that $PTCH_L$ plays a key role in embryogenesis.

The generation of mice in which one of the isoforms is nonfunctional may help clarify the roles of the alternative proteins in normal development and carcinogenesis. In this study, we focused on the usage of alternative first exons. However, some cell-surface receptors, such as CD44, undergo a complex, combinatorial splicing that determines the function of the gene products [39]. Although the major transcripts of human and mouse *PTCH* are ~8 kb long [18,20], we have identified rare transcripts lacking exons 4 and 5 (K.N. and T.M., unpublished data). Therefore, a comprehensive study of alternative pre-mRNA splicing throughout the gene using cost-effective and high-throughput methods, such as polymerase colony technology [40] or

exon junction microarrays [41], may shed light on the functional complexity of the *PTCH* gene and carcinogenesis with increased Shh pathway activity.

Materials and methods

Isolation of human *PTCH* isoforms and construction of plasmids

To obtain 5' ends of cDNA, RNA ligase-mediated 5'RACE was performed using the GeneRacer kit (Invitrogen) according to the manufacturer's directions. Random primers were used to reverse transcribe RNA. A reverse gene-specific primer was constructed in exon 2 to amplify the first-strand cDNA. The amplified cDNA was subcloned into pCR4-TOPO (Invitrogen) and sequenced. The expression plasmid encoding Myc-tagged *PTCH_L* (pMyc-Ptc1) was kindly provided by Dr. Jeffrey Ming. To make expression plasmids for *PTCH_M* and *PTCH_S*, a DNA fragment encoding the N-terminal region of *PTCH_L* was excised from pMyc-Ptc1 by digestion with *EcoRI* and replaced with the RT-PCR product encoding the N-terminal region of *PTCH_M* or *PTCH_S*, respectively. To make luciferase constructs, pGV-PTCH1, pGV-PTCH2, and pGV-PTCH3, fragments for the human *PTCH* promoter ranging from bp –1354 to –746, –4105 to –3808, and –8427 to –8032, respectively, relative to the reported transcription start site (GenBank Accession No. NM_000264) were subcloned into pGV-P2 (Wako Chemicals, Osaka, Japan). Mutated plasmids for these constructs were created by PCR-mediated mutagenesis as described previously [42]. The authenticity of all constructs was confirmed by DNA sequencing. The expression vector for FLAG-GLI1, pSRα-Flag-GLI1 [43], and the reporter vector, 8 × GLI-Luc [33], were kindly provided by Dr. Alexander L. Joyner and Hiroshi Sasaki, respectively.

Cell culture and transfections

The human embryonic kidney cell line 293T and mouse embryonal carcinoma cell line P19 were maintained in DMEM supplemented with 10% fetal calf serum (FCS), 50 U/ml penicillin, and 0.1 mg/ml streptomycin at 37°C in a humidified atmosphere of 5% CO₂. The human neuroblastoma line SH-SY5Y, oral squamous cell carcinoma lines HSC-2 and Ho-1-u-1, and lung squamous cell carcinoma line LK-2 (obtained from Cell Resource Center for Biomedical Research, Tohoku University, Japan) were maintained similarly except that RPMI 1640 medium was used. Cells were transfected with the indicated plasmids using Effectene reagent (Qiagen) and harvested at 16 h after the lipofection.

Analysis of *PTCH* isoform expression profiles

Human and mouse *PTCH* cDNA was amplified by RT-PCR using 0.5 µg of total RNA purified from a panel

of human tissues (Ambion and Clontech) or mouse embryos and primers 5'-CTGGGAGAAGACGGAGGAGGC-3' (exon 1a forward, human), 5'-CCCAGGAAATTAATAAAAGG-3' (exon 1a forward, mouse), 5'-GGACCGGGACTATCTGCACC-3' (exon 1b forward, human), 5'-GGACCGGGACTATCTGCACC-3' (exon 1b forward, mouse), 5'-CCTCTCCAGGAAAAGCAGCA-3' (exon 1c forward, human), 5'-GAGAAAAGCAGCAGACAAGTGAAGGTTG-3' (exon 1c forward, mouse), 5'-ATCCATGTGGCTGCCCTCTT-3' (exon 1d forward, human), 5'-ATCCTTGTGGCCGCCCTCTT-3' (exon 1d forward, mouse), 5'-TTCTCGGCGGG-GGTCCAGTT-3' (exon 1e forward, human), 5'-CCAGA-TGGACCACGGTTGCTGTAGATT-3' (exon 1e forward, mouse), 5'-CACAGCTCTCCACGTTGGT-3' (exon 2 reverse, human), and 5'-CACAGCTCTCCACGTTGGT-3' (exon 2 reverse, mouse). During the log phase of amplification (25–35 cycles depending on the templates), 1 µl of the PCR product was applied onto a DNA LabChip (Agilent Technologies) and loaded into an Agilent 2100 bioanalyzer according to the manufacturer's protocol. Data analysis was performed with Agilent 2100 bioanalyzer software. The expression of *PTCH* was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene or β-actin gene.

Western blotting

Immunoblot analysis was performed as described previously [44]. In brief, 30 µg of the cell lysate was subjected to SDS-PAGE and transferred onto a nitrocellulose membrane. The membrane was incubated with anti-c-Myc (Santa Cruz, 9E10) or anti-Flag (Sigma, M2) mouse monoclonal antibody followed by horseradish peroxidase-conjugated anti-mouse immunoglobulins (DAKO). The proteins were visualized using enhanced chemiluminescence immunoblotting detection reagents (Amersham).

Luciferase assay

293T or SH-SY5Y cells growing on six-well culture plates were cotransfected using Effectene reagent with various combinations of plasmids as indicated in the figure legends. Transfected cells were maintained in 0.5% FCS for 16 h and then harvested for the luciferase assay using the reagents and protocols provided by Promega or Wako chemicals.

Electrophoretic mobility shift assay

To obtain GST-GLI3 fusion protein, *Escherichia coli* strain BL21(DE3)pLysS (Novagen) was transformed with pGST-GLI3MF [45] (a gift from Dr. Shunsuke Ishii), which encodes the metal finger region of GLI3. The fusion protein was purified by affinity chromatography using glutathione-Sepharose 4B (Amersham Pharmacia Biotech) according to the manufacturer's instructions. The ³²P-labeled double-stranded oligonucleotide probes containing the sequence

for a consensus GLI-binding site (5'-TTGCCTACCTGGGTGGTCTCTCTACTT-3', 5'-CTCAGCCCTGACCCCAAGTCGAGCA-3', and 5'-GGCGCGGCAGACCACCCACGCCGAGGG-3') or a mutated sequence (5'-TTGCCTACCTAGTGGATCTCTCTACTT-3', 5'-CTCAGCCCTGATCCACTAAGTCGAGCA-3', and 5'-GGCGCGGCAGATCCACTACGCCGAGGG-3') (mutated nucleotides in *italic*) were incubated with the GST or GST-GLI3 fusion (200 ng). The reaction was performed in 10 μ l of binding buffer containing 4% glycerol, 1 mM $MgCl_2$, 0.5 mM EDTA, 50 mM NaCl, 10 mM Tris-HCl (pH 7.5), and 0.05 mg/ml poly(dI-dC) for 20 min at room temperature. For competition experiments, a 50-fold molar excess of unlabeled, double-stranded oligonucleotide, containing either a GLI site or a mutated GLI site as described above, was included in binding reactions. Samples were fractionated on a nondenaturing 6% polyacrylamide gel and visualized by autoradiography.

Apoptosis assay

293T cells were plated at 4×10^5 per well onto a six-well plate, grown for 16 h, transfected with plasmids for Myc-tagged PTCH or treated with cyclopamine (Toronto Research Chemicals, Inc.) (5 μ M final concentration), and then grown in DMEM with 0.5% FCS. After 24 h, relative DNA content was determined by flow cytometry as described previously [46]. Cells having a reduced DNA content (sub-G0/G1) were regarded as apoptotic.

Pulse-chase experiment

293 cells were plated at 8×10^5 cells per 60-mm plate, cultured for 24 h, and transfected with 1 μ g of PTCH expression plasmid using the Lipofectamine Plus reagent kit (Invitrogen) according to the manufacturer's instructions. Twenty-four hours after the transfection, cells were incubated with DMEM lacking methionine (-Met) for 30 min and then with 16.7 μ Ci/ml of L-[35 S]methionine with DMEM - Met for 2 h. Cells were washed three times with phosphate-buffered saline (PBS) and incubated with DMEM supplemented with 2 mM methionine for varying periods. At each time point, cells were scraped, washed with PBS, and lysed in 300 μ l of lysis buffer containing 150 mM NaCl, 1% Triton X-100, 10 mM Tris-HCl (pH 7.4), 5 mM EDTA, 1 mM PMSF, 18 μ g/ml aprotinin, 50 μ g/ml leupeptin, 1 mM benzamidin, and 0.7 μ g/ml pepstatin. The extracts were pelleted at 16,000g for 15 min at 4°C, and the supernatants (200 μ l) were immunoprecipitated for 16 h with 20 μ l of protein-A/G agarose (Santa Cruz) and 3 μ l of anti-c-Myc antibody. The immunoprecipitates were washed three times with 1 ml of lysis buffer, solubilized in 20 μ l of 1 \times Laemmli buffer by heating at 95°C for 5 min, and resolved on a 5–20% gradient polyacrylamide gel. Gels were dried, exposed, and analyzed using a FUJIX BAS2000 imaging analyzer (Fuji Film).

ChIP assay

ChIP assay was performed using the acetyl-histone H3 ChIP Assay Kit (Upstate Biotechnology), as recommended by the manufacturer, except that monoclonal anti-Flag (M2) and anti-Myc antibodies (9E10) were used in this study. 293T cells were plated at 1×10^6 cells per 10-cm dish, grown for 16 h, and then transfected with pFlag-GLI1 or pCI-Flag (encoding Flag-tag epitope). After 24 h, genomic DNA and protein were cross-linked by addition of formaldehyde (1% final concentration) directly to the culture medium and incubated for 10 min at 37°C. Cells were lysed in 1 ml of SDS lysis buffer containing 1% SDS, 10 mM EDTA, and 50 mM Tris-HCl (pH 8.1) and sonicated to generate 300- to 1000-bp DNA fragments. After centrifugation, the cleared supernatant was diluted 10-fold with ChIP dilution buffer and incubated with the specific antibody at 4°C for 16 h with rotation before incubation with protein A-Sepharose beads at 4°C for 1 h with rotation. Immune complexes were precipitated, washed, and eluted as recommended. DNA-protein cross-links were reversed by heating to 65°C for 2.5 h. DNA was phenol extracted, ethanol precipitated, and resuspended in 20 μ l of Tris-EDTA. We used 2.5 μ l of each sample as a template for PCR. PCR amplification was performed using primers that flank the putative GLI-response elements, 5'-AAAGGCTGGAGCTCCCGCCC-3' (GLI-BS1, forward) and 5'-TGCGCGCAAAGGCATCCAC-3' (GLI-BS1, reverse), or 5'-GGGCATGCATATTAAAGCCG-3' (GLI-BS2, forward) and 5'-CGAGCGCTATCTTAATCTCC-3' (GLI-BS2, reverse), or 5'-AGCGCCTGTTTACCCAGGAG-3' (GLI-BS3, forward) and 5'-GCTCCTCCGTCTTCTCCAG-3' (GLI-BS3, reverse).

Acknowledgments

We thank Mayu Yamazaki for technical support, Kayoko Saito for preparing the manuscript, and Drs. J. Ming (Children's Hospital of Philadelphia), A. Joyner (NYU Medical Center), S. Ishii (Tsukuba Life Science Center, RIKEN), and H. Sasaki (Center for Developmental Biology, RIKEN) for providing plasmids. This work was supported by Grants for Brain Research, Cancer Research, Genome Research, and Child Health and Development from the Ministry of Health, Labor, and Welfare, and a Grant-in-Aid for Scientific Research and the Budget for Nuclear Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

- [1] E. Belloni, et al., Identification of *Sonic hedgehog* as a candidate gene responsible for holoprosencephaly, *Nat. Genet.* 14 (1996) 353–356.
- [2] E. Rocssler, et al., Mutations in the human *Sonic hedgehog* gene cause holoprosencephaly, *Nat. Genet.* 14 (1996) 357–360.

- [3] C. Chiang, et al., Cyclopia and defective axial patterning in mice lacking *Sonic hedgehog* gene function, *Nature* 383 (1996) 407–431.
- [4] H. Hahn, et al., Mutations of the human homolog of *Drosophila patched* in the nevoid basal cell carcinoma syndrome, *Cell* 85 (1996) 841–851.
- [5] R.L. Johnson, et al., Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome, *Science* 272 (1996) 1668–1671.
- [6] A.B. Undén, et al., Mutations in the human homologue of *Drosophila patched* (*PTCH*) in basal cell carcinomas and the Gorlin syndrome: different *in vivo* mechanisms of *PTCH* inactivation, *Cancer Res.* 56 (1996) 4562–4565.
- [7] C. Raffel, et al., Sporadic medulloblastomas contain *PTCH* mutations, *Cancer Res.* 57 (1997) 842–845.
- [8] T. Pietsch, et al., Medulloblastomas of the desmoplastic variant carry mutations of the human homologue of *Drosophila patched*, *Cancer Res.* 57 (1997) 2085–2088.
- [9] R.J. Gorlin, Nevoid basal-cell carcinoma syndrome, *Medicine* 66 (1987) 98–113.
- [10] M.R. Gailani, et al., The role of the human homologue of *Drosophila patched* in sporadic basal cell carcinomas, *Nat. Genet.* 14 (1996) 78–81.
- [11] J. Xie, et al., Activating *Smoothened* mutations in sporadic basal-cell carcinoma, *Nature* 391 (1998) 90–92.
- [12] S.P. Thayer, et al., Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis, *Nature* 425 (2003) 851–856.
- [13] D.M. Berman, et al., Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours, *Nature* 425 (2003) 846–851.
- [14] D.N. Watkins, et al., Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer, *Nature* 422 (2003) 313–317.
- [15] P.W. Ingham, A.P. McMahon, Hedgehog signaling in animal development: paradigms and principles, *Genes Dev.* 15 (2001) 3059–3087.
- [16] D.M. Berman, Medulloblastoma growth inhibition by hedgehog pathway blockade, *Science* 297 (2002) 1559–1561.
- [17] P. Kogerman, et al., Alternative first exons of *PTCH1* are differentially regulated *in vivo* and may confer different functions to the *PTCH1* protein, *Oncogene* 21 (2002) 6007–6016.
- [18] L.V. Goodrich, R.L. Johnson, L. Milenkovic, J.A. McMahon, M.P. Scott, Conservation of the hedgehog/patched signaling pathway from flies to mice: induction of a mouse *patched* gene by Hedgehog, *Genes Dev.* 10 (1996) 301–312.
- [19] Q. Wu, A.R. Krainer, AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes, *Mol. Cell Biol.* 19 (1999) 3225–3236.
- [20] H. Hahn, et al., A mammalian *patched* homolog is expressed in target tissues of *sonic hedgehog* and maps to a region associated with developmental abnormalities, *J. Biol. Chem.* 271 (1996) 12125–12128.
- [21] K.W. Kinzler, B. Vogelstein, The *GLI* gene encodes a nuclear protein which binds specific sequences in the human genome, *Mol. Cell Biol.* 10 (1990) 634–642.
- [22] C. Alexandre, A. Jacinto, P.W. Ingham, Transcriptional activation of *hedgehog* target genes in *Drosophila* is mediated directly by the cubitus interruptus protein, a member of the *GLI* family of zinc finger DNA-binding proteins, *Genes Dev.* 10 (1996) 2003–2013.
- [23] M. Agren, P. Kogerman, M.I. Kleman, M. Wessling, R. Toftgård, Expression of the *PTCH1* tumor suppressor gene is regulated by alternative promoters and a single functional *Gli*-binding site, *Gene* 330 (2004) 101–114.
- [24] E.A. Barnes, M. Kong, V. Ollendorff, D.J. Donoghue, Patched1 interacts with cyclin B1 to regulate cell cycle progression, *EMBO J.* 20 (2001) 2214–2223.
- [25] C. Thibert, et al., Inhibition of neuroepithelial patched-induced apoptosis by sonic hedgehog, *Science* 301 (2003) 843–846.
- [26] J.K. Chen, J. Taipale, M.K. Cooper, P.A. Beachy, Inhibition of Hedgehog signaling by direct binding of cyclopamine to *Smoothened*, *Genes Dev.* 16 (2002) 2743–2748.
- [27] E.C. Bailey, L. Milenkovic, M.P. Scott, J.F. Collawn, R.L. Johnson, Several *PATCHED1* missense mutations display activity in *patched1*-deficient fibroblasts, *J. Biol. Chem.* 277 (2002) 33632–33640.
- [28] V. Marigo, R.A. Davey, Y. Zuo, J.M. Cunningham, C.J. Tabin, Biochemical evidence that *patched* is the Hedgehog receptor, *Nature* 384 (1996) 176–179.
- [29] Z. Kan, E.C. Rouchka, W.R. Gish, D.J. States, Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, *Genome Res.* 11 (2001) 889–900.
- [30] P.D. Stenson, et al., Human Gene Mutation Database (HGMD): 2003 update, *Hum. Mutat.* 21 (2003) 577–581.
- [31] J.J. Fuentes, M.A. Pritchard, X. Estivill, Genomic organization, alternative splicing, and expression patterns of the *DSCR1* (Down syndrome candidate region 1) gene, *Genomics* 44 (1997) 358–361.
- [32] Y. Wang, et al., RNA diversity has profound effects on the translation of neuronal nitric oxide synthase, *Proc. Natl. Acad. Sci. U.S.A.* 96 (1999) 12150–12155.
- [33] H. Sasaki, C. Hui, M. Nakafuku, H. Kondoh, A binding site for *Gli* proteins is essential for *HNF-3 β* floor plate enhancer activity in transgenics and can respond to *Shh* *in vitro*, *Development* 124 (1997) 1313–1322.
- [34] K. Fujii, et al., Mutations in the human homologue of *Drosophila patched* in Japanese nevoid basal cell carcinoma syndrome patients, *Hum. Mutat.* 21 (2003) 451–452.
- [35] A. Chidambaram, et al., Mutations in the human homologue of the *Drosophila patched* gene in Caucasian and African-American nevoid basal cell carcinoma syndrome patients, *Cancer Res.* 56 (1996) 4599–4601.
- [36] C. Wicking, et al., Most germ-line mutations in the nevoid basal cell carcinoma syndrome lead to a premature termination of the *PATCHED* protein, and no genotype-phenotype correlations are evident, *Am. J. Hum. Genet.* 60 (1997) 21–26.
- [37] I. Smyth, et al., Isolation and characterization of human *patched 2* (*PTCH2*), a putative tumour suppressor gene in basal cell carcinoma and medulloblastoma on chromosome 1p32, *Hum. Mol. Genet.* 8 (1999) 291–297.
- [38] P.G. Zaphiropoulos, A.B. Undén, F. Rahnama, R.E. Hollingsworth, R. Toftgård, *PTCH2*, a novel human *patched* gene, undergoing alternative splicing and up-regulated in basal cell carcinomas, *Cancer Res.* 59 (1999) 787–792.
- [39] J. Lesley, R. Hyan, CD44 structure and function, *Front. Biosci.* 3 (1998) D616–D630.
- [40] J. Zhu, J. Shendure, R.D. Mitra, G.M. Church, Single molecule profiling of alternative pre-mRNA splicing, *Science* 301 (2003) 836–838.
- [41] J.M. Johnson, et al., Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays, *Science* 302 (2003) 2141–2144.
- [42] Y. Imai, Y. Matsushima, T. Sugimura, M. Terada, A simple and rapid method for generating a deletion by PCR, *Nucleic Acids Res.* 19 (1991) 2785.
- [43] H.L. Park, et al., Mouse *Gli1* mutants are viable but have defects in SHH signaling in combination with a *Gli2* mutation, *Development* 127 (2000) 1593–1605.
- [44] T. Miyashita, Y. Okamura-Oho, Y. Mito, S. Nagafuchi, M. Yamada, Dentatorubral pallidoluysian atrophy (DRPLA) protein is cleaved by caspase-3 during apoptosis, *J. Biol. Chem.* 272 (1997) 29238–29242.
- [45] P. Dai, et al., Sonic Hedgehog-induced activation of the *Gli1* promoter is mediated by *GLI3*, *J. Biol. Chem.* 274 (1999) 8143–8152.
- [46] K. Fujii, et al., γ -Irradiation deregulates cell cycle control and apoptosis in nevoid basal cell carcinoma syndrome-derived cells, *Jpn. J. Cancer Res.* 90 (1999) 1351–1357.
- [47] B.C. Schaefer, Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends, *Anal. Biochem.* 227 (1995) 255–273.

An Autoantibody-Mediated Immune Response to Calreticulin Isoforms in Pancreatic Cancer

Su-Hyung Hong,¹ David E. Misek,¹ Hong Wang,¹ Eric Puravs,¹ Thomas J. Giordano,² Joel K. Greenson,² Dean E. Brenner,³ Diane M. Simeone,⁴ Craig D. Logsdon,⁵ and Samir M. Hanash¹

Departments of ¹Pediatrics, ²Pathology, ³Internal Medicine, ⁴Surgery, and ⁵Physiology, University of Michigan Medical School, Ann Arbor, Michigan

ABSTRACT

The identification of circulating tumor antigens or their related autoantibodies provides a means for early cancer diagnosis as well as leads for therapy. We have used a proteomic approach to identify proteins that commonly induce a humoral response in pancreatic cancer. Aliquots of solubilized proteins from a pancreatic cancer cell line (Panc-1) were subjected to two-dimensional PAGE, followed by Western blot analysis in which sera of individual patients were tested for primary antibodies. Sera from 36 newly diagnosed patients with pancreatic cancer, 18 patients with chronic pancreatitis, 33 patients with other cancers, and 15 healthy subjects were analyzed. Autoantibodies were detected against either one or two calreticulin isoforms identified by mass spectrometry in sera from 21 of 36 patients with pancreatic cancer. One of 18 chronic pancreatitis patients and 1 of 15 healthy controls demonstrated autoantibodies to calreticulin isoform 1; none demonstrated autoantibodies to isoform 2. None of the sera from patients with colon cancer exhibited reactivity against either of these two proteins. One of 14 sera from lung adenocarcinoma patients demonstrated autoantibodies to calreticulin isoform 1; 2 of 14 demonstrated autoantibodies to isoform 2. Immunohistochemical analysis of calreticulin in pancreatic/ampullary tumor tissue arrays using an isoform nonspecific antibody revealed diffuse and consistent cytoplasmic staining in the neoplastic epithelial cells of the pancreatic and ampullary adenocarcinomas. The detection of autoantibodies to calreticulin isoforms may have utility for the early diagnosis of pancreatic cancer.

INTRODUCTION

There is, at present, much interest in identifying markers for the early detection of pancreatic cancer. We have implemented a proteomics-based approach to identify tumor markers based on their occurrence as tumor antigens that elicit a humoral response during tumorigenesis. The humoral immune response to cancer in humans has been well demonstrated by identification of autoantibodies to a number of different intracellular and surface antigens in patients with various tumor types (1-4).

Pancreatic cancer has the worst prognosis of all cancers, with a 5-year survival rate of <3%, accounting for the fourth largest number of cancer deaths in the United States (5). It occurs with a frequency of around 9 patients/100,000 individuals, making it the 11th most common cancer in the United States. The poor prognosis for pancreatic cancer is due, in part, to lack of early diagnosis. There is currently no effective biomarker-based strategy useful for the early detection of pancreatic cancer or even to differentiate between pancreatic adenocarcinoma and chronic pancreatitis. In pancreatic cancer, autoimmunity has been shown against several proteins, including MUC1 (6),

p53 (7), and Rad51 (8) proteins. MUC1 is a transmembrane glycoprotein involved in cell-cell and cell-extracellular matrix interactions, and MUC1 autoantibodies have been observed in sera from patients with a variety of different tumors (9). In pancreatic cancer, the presence of MUC1 IgG autoantibodies has been shown to be associated with a favorable prognosis (6). The presence of p53 autoantibodies has been observed in 18.2% of patients with pancreatic cancer. However, p53 autoantibodies were also found in 5.3% of patients with acute pancreatitis and 12.1% of patients with chronic pancreatitis, thus the humoral response to p53 was not specific to malignancy. The recombination factor Rad51 is highly expressed in pancreatic adenocarcinoma (10), and Rad51 autoantibodies have been observed in 7% of patients with pancreatic cancer.

It is not clear why only a subset of patients with a particular tumor type develop a humoral response to a particular antigen. Immunogenicity may depend on the level of expression, posttranslational modification, or other types of protein processing, the extent of which may be variable among tumors of a similar type. Other factors that may influence the immune response include variability among tumors and individuals in MHC molecules and in antigen presentation. A large number of autoantibodies have been identified in different tumor types, but in most cases, they occur in less than 50% of sera of patients. Therefore, they are not effective individually for the early detection of cancer. Thus, the development of panels of such autoantibodies directed against a variety of tumor antigens may be effective (11).

The identification of panels of tumor antigens that elicit an immune response may have utility in early cancer diagnosis, in establishing prognosis, and in immunotherapy against the disease. Several approaches are currently available for the identification of tumor antigens. In contrast to identification of tumor antigens based on analysis of recombinant proteins, the proteomic-based approach for the identification of tumor antigens that we have used allows for the identification of autoantibodies to proteins as they occurred in their natural states, in lysates prepared from tumors and tumor cell lines. This technology may uncover antigenicity associated with aberrant posttranslational modification of tumor cell proteins. The goal of this study was to implement a proteomic approach for the identification of tumor antigens that elicit a humoral response in pancreatic cancer cell line, Panc-1. To this end, we have used two-dimensional PAGE to simultaneously separate individual cellular proteins from the Panc-1 cell line. The separated proteins were transferred onto polyvinylidene difluoride membranes. Sera from cancer patients were screened individually for antibodies that reacted against the separated proteins by Western blot analysis. Proteins specifically reacting with sera from cancer patients were identified by mass spectrometry. We have identified two calreticulin isoforms as proteins that commonly elicit an antibody response in pancreatic cancer.

MATERIALS AND METHODS

Sera and Cell Lines. Serum and tumor tissue were obtained at the time of diagnosis following informed consent. The experimental protocol was approved by The University of Michigan Institutional Review Board. Sera were

Received 1/10/04; revised 5/4/04; accepted 5/20/04.

Grant support: National Cancer Institute Grant U01 CA-84982 (S. Hanash), Michigan Economic Development Corporation-Life Sciences Corridor Fund Grant 03-622 (C. Logsdon), and Tissue Core of the University of Michigan Comprehensive Cancer Center Grant CA46952.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: S. Hong is currently at the Department of Dental Microbiology, School of Dentistry, Kyungpook National University, Jung-Gu, Daegu, South Korea.

Requests for reprints: David E. Misek, Department of Pediatrics, University of Michigan, Room A520 MSRB-1, 1150 West Medical Center Drive, Ann Arbor, MI 48109-0656. Phone: (734) 763-9311; E-mail: dmisek@umich.edu.

CALRETICULIN AS A BIOMARKER IN PANCREATIC CANCER

obtained from 36 patients with pancreatic cancer (all of advanced stage). Sera from 18 patients with chronic pancreatitis, from 15 healthy individuals, and from 33 patients with other cancers (14 with lung cancer and 19 with colon cancer) were used as controls. All subjects that donated sera for this study were between 57 and 74 years of age. The human cancer cell lines used in this study were all individually cultured in DMEM supplemented with 10% fetal bovine serum, 100 units/ml penicillin, and 100 units/ml streptomycin (Invitrogen, Carlsbad, CA).

Two-Dimensional PAGE and Western Blot Analysis. After excision, the tumor tissue was immediately frozen at -80°C , after which an aliquot was lysed in solubilization buffer [8 M urea (Bio-Rad), 2% NP40, 2% carrier ampholytes (pH 4–8; Gallard/Schlessinger, Carle Place, NY), 2% β -mercaptoethanol, and 10 mM phenylmethylsulfonyl fluoride] and stored at -80°C until use. Cultured Panc-1 pancreatic adenocarcinoma cells were harvested in 300 μl of solubilization buffer by using a cell scraper and stored at -80°C until use. Proteins derived from the extracts of either cultured cells or solid tumors were separated into two dimensions as described previously (12). In brief, solubilized proteins were applied onto isoelectric focusing gels. Isoelectric focusing was performed using pH 4–8 carrier ampholytes at 700 V for 16 h, followed by 1000 V for an additional 2 h. The first-dimension gel was loaded onto the second-dimension gel, after equilibration in 125 mM Tris (pH 6.8), 10% glycerol, 2% SDS, 1% DTT, and bromophenol blue. For the second-dimension separation, a gradient of 11–14% acrylamide (Crescent Chemical, Huppauge, NY) was used. Proteins were transferred to an Immobilon-P polyvinylidene difluoride membrane (Millipore, Bedford, MA) or visualized by silver staining of the gels.

Western Blotting. After transfer, membranes were incubated with a blocking buffer consisting of 10 mM Tris-HCl (pH 7.5), 50 mM NaCl, 1.8% nonfat dry milk, and 0.01% Tween 20 for 2 h. The membranes were incubated for 1 h at room temperature with serum obtained from either patients or healthy individuals as a source of primary antibody at a 1:100 dilution. After three washes with washing buffer (Tris-buffered saline containing 0.01% Tween 20), the membranes were incubated with horseradish peroxidase-conjugated sheep anti-human (Amersham Biosciences, Piscataway, NJ) IgG antibodies at a dilution of 1:1000 for 1 h at room temperature. Immunodetection was accomplished by enhanced chemiluminescence (Amersham Biosciences) followed by autoradiography on Hyperfilm MP (Amersham Biosciences).

Calreticulin Detection by Western Blotting. A rabbit anticalreticulin polyclonal antibody (Affinity Bioreagents, Golden, CO) was used at a 1:1000 dilution for Western blotting and was processed as for incubations with patient sera, with a horseradish peroxidase-conjugated anti-rabbit IgG (Amersham Biosciences) as the secondary antibody.

In-Gel Enzyme Digestion and Mass Spectrometry. For protein identification by mass spectrometry, two-dimensional gels were stained by a modified silver-staining method, and excised proteins were destained for 5 min in 15 mM potassium ferricyanide and 50 mM sodium thiosulfate as described previously (13). After three washes with water, the gel pieces were dehydrated in 100% acetonitrile for 5 min and then dried. Digestion was performed by the addition of 100 ng of trypsin (Promega, Madison, WI) in 200 nM ammonium bicarbonate. After enzymatic digestion overnight at 37°C , the peptides were extracted twice with 50 μl of 60% acetonitrile/1% trifluoroacetic acid. After removal of acetonitrile by centrifugation in a vacuum centrifuge, the peptides were concentrated by using pipette tips C18 (Millipore) and identified by nanoflow capillary liquid chromatography coupled with electrospray quadrupole-time of flight tandem mass spectrometry in the quadrupole-time of flight micro (MikroMass, Manchester, United Kingdom). The acquired spectra were processed and searched against a nonredundant SwissProt protein sequence database using proteinLynx Global Server.*

RNA Isolation. Samples of normal pancreas were taken from organ donors provided by the Michigan Transplantation Society (five) or from areas outside regions of pathology in surgically resected pancreata (two). All of the pancreatic cancers were of advanced stage. All samples were processed in a similar manner. Frozen samples were embedded in OCT-freezing media (Miles Scientific, Naperville, IL) and cryotome sectioned (5 μm), and routine H&E stains were evaluated by a surgical pathologist. Areas of relatively pure tumor (at least 70% tumor cells) or normal tissue were microdissected, and these

Table 1. Calreticulin autoantibodies in patients sera

Subjects	Number of subjects	Isoform 1 positive	Isoform 2 positive	Isoform 1 or 2 positive
Pancreatic cancer	36	15 (41.7%)	16 (44.4%)	21 (58.3%)
Chronic pancreatitis	18	1 (5.6%)	0 (0%)	1 (5.6%)
Control groups				
Lung cancer	14	1 (7.1%)	2 (14.3%)	2 (14.3%)
Colon cancer	19	0 (0%)	0 (0%)	0 (0%)
Healthy subjects	15	1 (6.7%)	0 (0%)	1 (6.7%)

areas were selected for RNA isolation. All grades of differentiation were exhibited by the tumors.

Isolates of human tumor tissue and human tumor cell lines were homogenized in the presence of TRIzol reagent (Invitrogen), and total cellular RNA was purified according to the manufacturer's procedures. RNA samples were further purified using acid phenol extraction and RNeasy spin columns (Qiagen, Valencia, CA). RNA quality was assessed by 1% agarose gel electrophoresis in the presence of ethidium bromide.

Gene Expression Profiling and Statistical Analysis. This study used commercially available high-density oligonucleotide microarrays (U133A; Affymetrix, Santa Clara, CA). Hybridization, scanning, and image analysis of the arrays were performed according to manufacturer's protocols and as described previously (14, 15). The U133A array consists of 22,283 probe sets, each representing a transcript. Each probe set typically consists of 11 perfectly complementary 25-base-long probes (PM) as well as 11 mismatch probes (MM) that are identical except for an altered central base. A normal pancreas sample was selected as the standard, and probe pairs for which PM/MM ≤ 100 on the standard were excluded from additional analysis. The average of the middle 50% of the PM/MM differences was used as the expression measure for each probe set. A quantile normalization procedure was used to adjust for differences in the probe intensity distribution across different chips. In brief, we applied a monotone linear spline to each chip that mapped quantiles 0.01 up to 0.99 (in increments of 0.01) exactly to the corresponding quantiles of the standard. For statistical analysis, we first transformed each normalized probe-set expression value, x , to $\log[100 + \max(x + 100, 0)]$, which we found stabilized the within group variances between high- and low-expression probe sets. To compare normal and tumor samples, we performed a one-way ANOVA, modeling the log-transformed values for each probe set as having separate means for each group. We calculated fold changes between groups of samples by first replacing mean expression values < 100 units by 100 to avoid negative values or spuriously large fold changes. Code to perform these computations is freely available.⁷

Determination of Calreticulin mRNA Levels Using Real-Time PCR. Five pancreatic, four lung, three colon and two ovarian cancer cell lines were used to compare the mRNA expression level of calreticulin. Expression levels were normalized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA expression. Oligonucleotide primers and TaqMan probes were designed using the Light Cycler Probe Design Software (Roche Applied Science). Forward and reverse primers for human calreticulin were 5'-CGCC-ATCTGCTATCC-3' and 5'-CATAAAAGCGTGCATCT-3', respectively (Applied Biosystems). The nucleotide sequence of the forward and reverse primers for GAPDH were 5'-GAAGGTGAAGGTGGAGTC-3' and 5'-GAAGATGGTGATGGGATTTC-3', respectively (Applied Biosystems).

The first-strand cDNA was synthesized with SuperScript First-Strand Synthesis System for reverse transcription-PCR according to the manufacturer's instructions (Invitrogen). Quantitative PCR reaction was carried out in 96-well optical reaction plates using cDNA derived from 50 ng of total RNA for each sample in a volume of 25 μl . PCR was performed on the ABI Prism 7700 Sequence Detector (Applied Biosystems). The cycling conditions were 10 min at 95°C followed by 55 cycles at 95°C for 30 s, 60°C for 45 s, and 72°C for 45 s.

To control for the variation in the amount of starting RNA among samples, we performed amplification of GAPDH mRNA as an internal reference against which other RNA values were normalized. Additionally, the real-time PCR

* Internet address: www.mikromass.co.uk.

⁷ Internet address: http://dub.pict.umich.edu:2000/ourimage/pub/index.html.

CALRETICULIN AS A BIOMARKER IN PANCREATIC CANCER

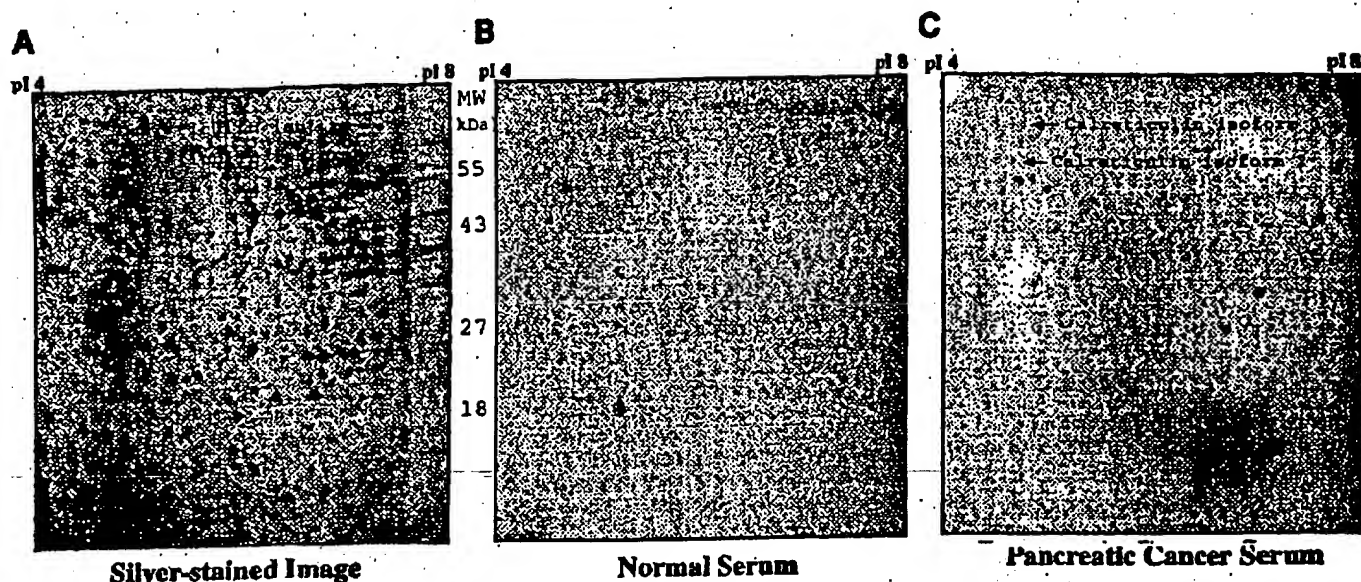


Fig. 1. A silver-stained image of the Panc-1 pancreatic tumor cell line (A) compared with a Western blot of the Panc-1 cell line with normal serum (B) and serum from a patient with pancreatic cancer (C).

products were purified by QIAQuick Gel Extraction kit (Qiagen) and subjected to DNA sequencing to verify the identity of the real-time PCR products.

Pancreas/Ampullary Tumor Tissue Array and Immunohistochemistry. A tissue array containing triplicates of 4 normal pancreas, 12 nonpancreas normal tissues, 47 pancreatic adenocarcinomas, 31 ampullary adenocarcinomas, and 2 large cell anaplastic carcinomas was constructed as described previously (9). The cases were randomly selected from the University of Michigan Pathology archives. Immunohistochemistry for calreticulin was performed using the same rabbit polyclonal antibody (30 min incubation at room temperature) at 1:200 using citrate buffer (pH 6.0) and microwave antigen retrieval (10 min) and the Dako automated instrument (Dako Cytomation, Carpinteria, CA). Primary antibody was detected using the Dako Envision kit.

RESULTS

Pancreatic Tumor Proteins Recognized Specifically by Sera from Newly Diagnosed Patients with Pancreatic Cancer. Panc-1 pancreatic tumor cell line proteins were separated by two-dimensional PAGE and transferred onto Immobilon-P polyvinylidene difluoride membranes. Sera obtained from 36 newly diagnosed patients with pancreatic cancer, from 18 patients with chronic pancreatitis, from 33 patients with other types of cancers, and from 15 healthy donors were screened individually for the presence of antibodies to Panc-1 pancreatic tumor cell line proteins (Table 1). Each membrane was treated

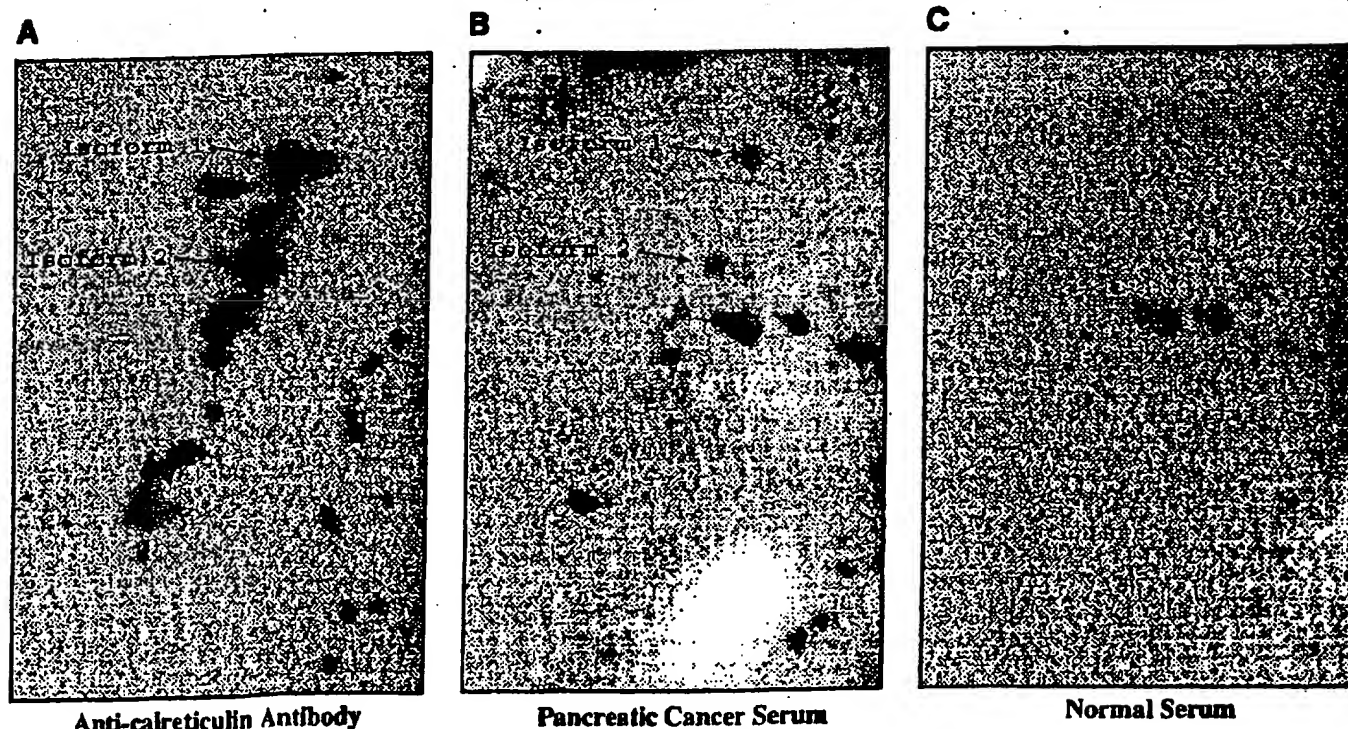
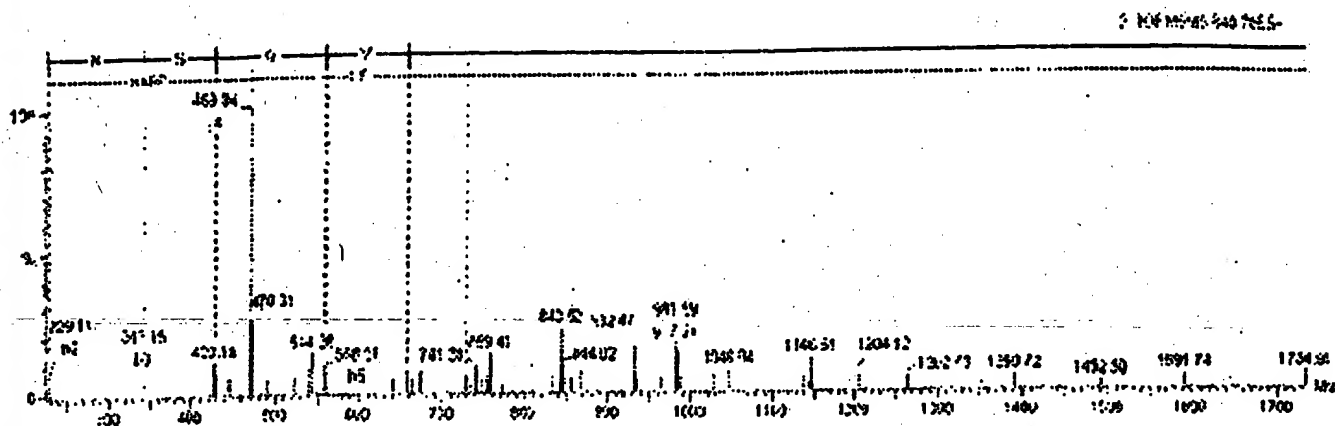


Fig. 2. Western blot analysis of calreticulin with a polyclonal anticalreticulin antibody (A) and sera from a pancreatic cancer patient (B) and from a healthy individual (C).

5506

CALRETICULIN AS A BIOMARKER IN PANCREATIC CANCER

m/z	MU	Delta	Score	Start	End	Sequence
840.4000	2518.1860	0.01	12.74	186	207	(K) IDNSQ VESGS LEDDW DFLPP KK(I)
1196.0524	2390.0910	-0.02	5.73	186	206	(K) IDNSQ VESGS LEDDW DFLPP K(K)



1	MLLSVPLLLG	LIGLAVAEPA	VYFKEQFLDG	DGWTSRWIES	KHKSDPGKFV	LSSGKFYGD	EKDRLQTSQ
71	DARFYALSAS	FEPFSNKGQT	LVVQFTVKHE	QNIDCGGGYV	KLFPNSLDQT	DMHGDSEYNI	MFGPDICGPG
141	TKKVHVFYNY	KGKNVLINKD	IRCKDDEFTH	LYTLIVRPDN	TYEVKIDNSQ	VESGSLEDW	DFLPPKRIKD
211	PDASKPEDWD	ERIKDDPTD	SKPEDWDKPE	HIFDPDAKKP	EDWDEEMDGE	NEPPVIQNP	EYGEWKPRCI
281	DNPDYKGTWI	HPEIDNPEYS	PDPSTIAYDN	FGVLGLDLNQ	VKSGTIFDNP	LITNDEAYAE	EFGMETWGV
351	KAAEKQMKDK	QDEEQRLKEE	EDDKRKEEE	EAEDKFDD	KDEDEDEED	KEEDEEDVP	GQAKDEL

Fig. 3. Tandem mass spectrometry identification of calreticulin isoform 1. The tandem mass spectrometry spectrum of calreticulin isoform 1 (obtained after trypsin digestion) is shown by analysis with electrospray quadrupole-time of flight, coupled with nanoflow capillary high-performance liquid chromatography. The precursor ion shown in the figure is m/z 840.4000, and the resultant peaks were searched against the nonredundant SwissProt protein sequence database using the ProteinLynx global server.

with one serum sample as the primary antibody and with sheep antihuman IgG as the secondary antibody. In general, most pancreatic patient sera reacted against multiple spots (Fig. 1; Fig. 2). Some of the reactive protein spots were observed in the control sera and thus were considered to represent nonspecific reactivity. The reactive proteins most commonly observed with pancreatic cancer patient sera but not with noncancer controls included two proteins (spots 1 and 2) with an estimated molecular mass of 55–60 kDa and an isoelectric point of 4.4. These two proteins frequently showed concordant reactivity with the same sera suggesting, given their close proximity in two-dimensional gels, that they represented isoforms of the same protein. The protein from spot 1 showed reactivity with sera from 17 of 36 patients with pancreatic cancer (47.2%), with sera from 1 of 18 patients with chronic pancreatitis (5.6%) and with sera from 1 of 15 healthy donors (6.7%). The protein from spot 2 showed reactivity in 16 of 36 patients with pancreatic cancer (44.4%), in 0 of 18 patients with chronic pancreatitis and in 0 of 15 healthy donors (Table 1). The number of pancreatic cancer patients' sera that showed reactivity with one or both spots was 21 of 36 (58.3%). Reactivity directed against the protein in spot 1 was found in sera from 1 of 14 patients with lung cancer; reactivity directed against the protein in spot 2 was found in 2 of 14 lung cancer patients. None of the sera from 19 colon cancer patients exhibited reactivity against the protein in either spot (1 or 2).

Identification of the Reactive Proteins as Isoforms of Calreticulin. The proteins of interest were extracted from the gels after two-dimensional PAGE and silver staining. The proteins were digested with trypsin, and the resulting peptides were analyzed by

electrospray quadrupole-time of flight tandem mass spectrometry. The acquired spectra were processed and searched against a nonredundant SwissProt protein sequence database using ProteinLynx Global Server.⁶ The two proteins were identified (Fig. 3) as being isoforms of calreticulin, (SwissProt accession no. P27797). Identity with calreticulin was confirmed with two-dimensional Western blotting using Panc-1 whole-cell extracts and an anticacreticulin rabbit polyclonal antibody.

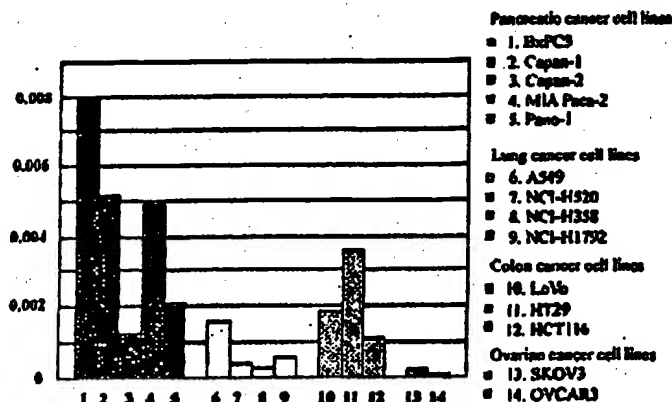


Fig. 4. Calreticulin mRNA levels measured by real-time PCR in pancreatic, lung, colon, and ovarian tumor cell lines expressed as calreticulin:GAPDH ratio, as described in "Materials and Methods." Each bar represents the mean \pm SE of five independent experiments.

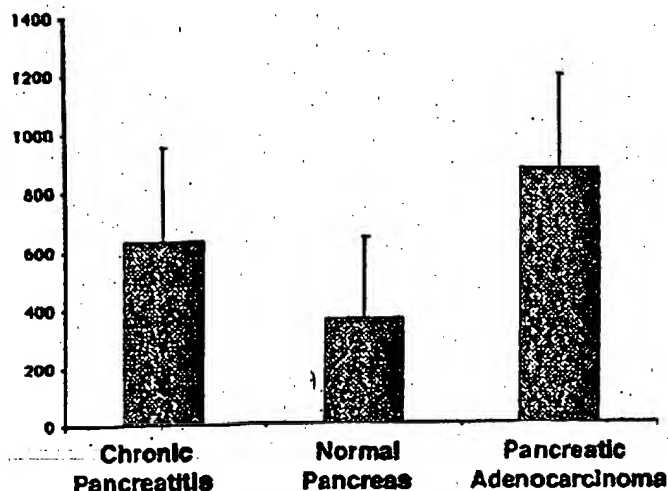


Fig. 5. DNA microarray data of calreticulin in chronic pancreatitis (4), normal pancreas (7), and pancreatic adenocarcinoma (8). Each bar represents the mean \pm SD

Role of Glycosylation in Calreticulin Antigenicity. We sought to determine whether calreticulin glycosylation contributed to immunogenicity. Solubilized proteins from the Panc-1 pancreatic tumor cell line were subjected to *N*-deglycosylation by a combination of endoglycosidase F, endo- α -*N*-acetylgalactosaminidase, and α -2-3,6,8,9-neuraminidase. The resulting products were separated by SDS electrophoresis and analyzed by Western blotting. Although the deglycosylated positive control revealed a demonstrable mobility shift by SDS-PAGE, the deglycosylating enzyme treatment did not result in any mobility shifts of calreticulin. Thus, endoglycosidase F-sensitive glycosylation does not appear to play a role in the observed immunogenicity of the calreticulin isoforms (data not shown).

mRNA Expression of Calreticulin. To examine whether the immunogenicity of calreticulin in pancreatic cancer could be due to elevated transcriptional mechanisms, the expression of calreticulin

mRNA was examined in different cell lines and tumor tissues. To examine calreticulin expression in all cell lines, including five pancreatic tumor cell lines, four lung tumor cell lines, three colon tumor cell lines, and two ovarian tumor cell lines, we performed real time-PCR using the expression level of GAPDH as an internal control. After normalization, the calreticulin:GAPDH ratio was calculated from each cell line (Fig. 4). In general, we found that the level of mRNA expression in the pancreatic tumor cell lines was significantly higher than the other cell lines examined, suggesting that overexpression of calreticulin may be a possible contributing factor in its immunogenicity. Therefore, we examined calreticulin expression in eight pancreatic adenocarcinomas, in four samples of chronic pancreatitis, and in seven samples of normal pancreas by microarray analysis (Fig. 5). The expression of calreticulin mRNA was approximately 2-fold higher in pancreatic tumors as compared with normal pancreas ($P = 0.006$). It is important to note, however, that the pancreatic adenocarcinomas were microdissected and are derived from ductal epithelium. Because the normal pancreas is primarily acinar, it may be that the difference in gene expression noted in the pancreas tumors is entirely related to the differences in the epithelium analyzed rather than any differences that arose in the tumor.

Analysis of Calreticulin Expression by Two-Dimensional PAGE. We hypothesized that there might be changes in the levels of calreticulin total protein or isoforms that could lead to antigenicity in pancreatic cancer. Using two-dimensional PAGE, we examined the expression of calreticulin isoforms 1 and 2 in a variety of tissues and tumor types. All calreticulin isoforms were present in different cell lines, including 6 pancreatic tumor cell lines, 4 lung tumor cell lines, 9 colon tumor cell lines, and 33 ovarian tumor cell lines, at similar expression levels. A similar pattern of expression was also observed in 6 pancreatic tumors, 38 lung tumors, 7 colon tumors, and 25 ovarian tumors (Fig. 6). All isoforms were also observed in a variety of normal tissues and in gastric, esophageal, and brain tumors (data not shown). These results suggest that all calreticulin isoforms, including isoforms 1 and 2, were ubiquitously expressed and that the

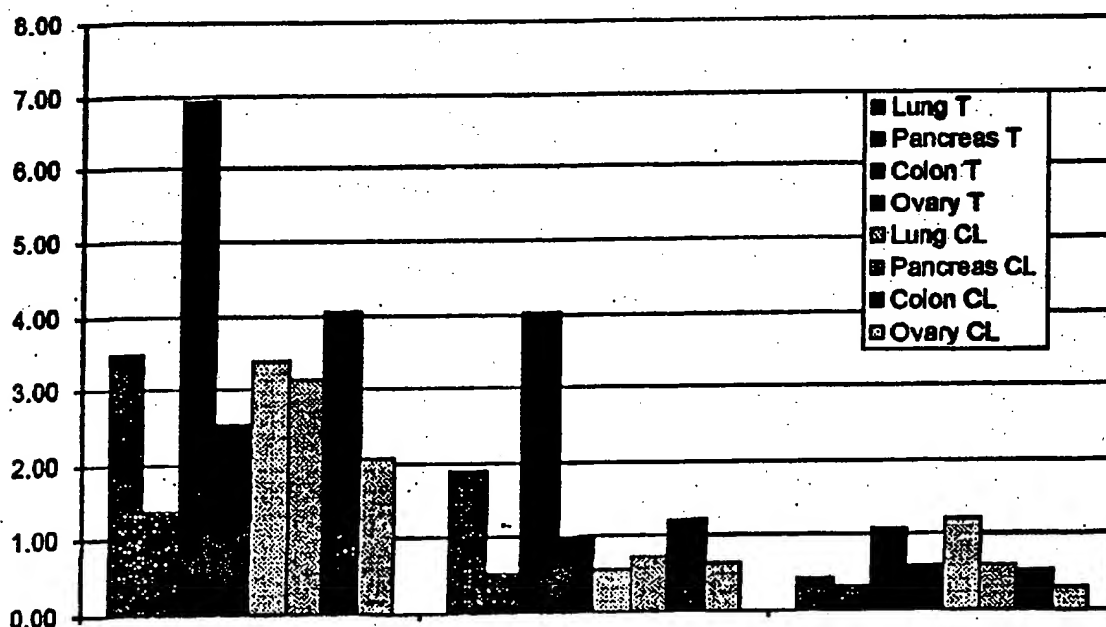


Fig. 6. Calreticulin protein levels measured in human tumors (T) and tumor cell lines (CL). Two-dimensional gels were prepared using solubilized proteins from a variety of human tumors and tumor cell lines, as described in "Materials and Methods." Background-corrected integrated intensity (volume) was measured (Viasage software; Genomic Solutions, Ann Arbor, MI) for total calreticulin, calreticulin isoform 1, and calreticulin isoform 2 and was normalized to the values obtained from the average of two unidentified invariant spots. Bars represent the average intensities for 38 lung tumors, 6 pancreatic tumors, 7 colon tumors, 25 ovarian tumors, 4 lung cell lines, 6 pancreatic cell lines, 9 colon cell lines, and 33 ovarian cell lines.

CALRETICULIN AS A BIOMARKER IN PANCREATIC CANCER

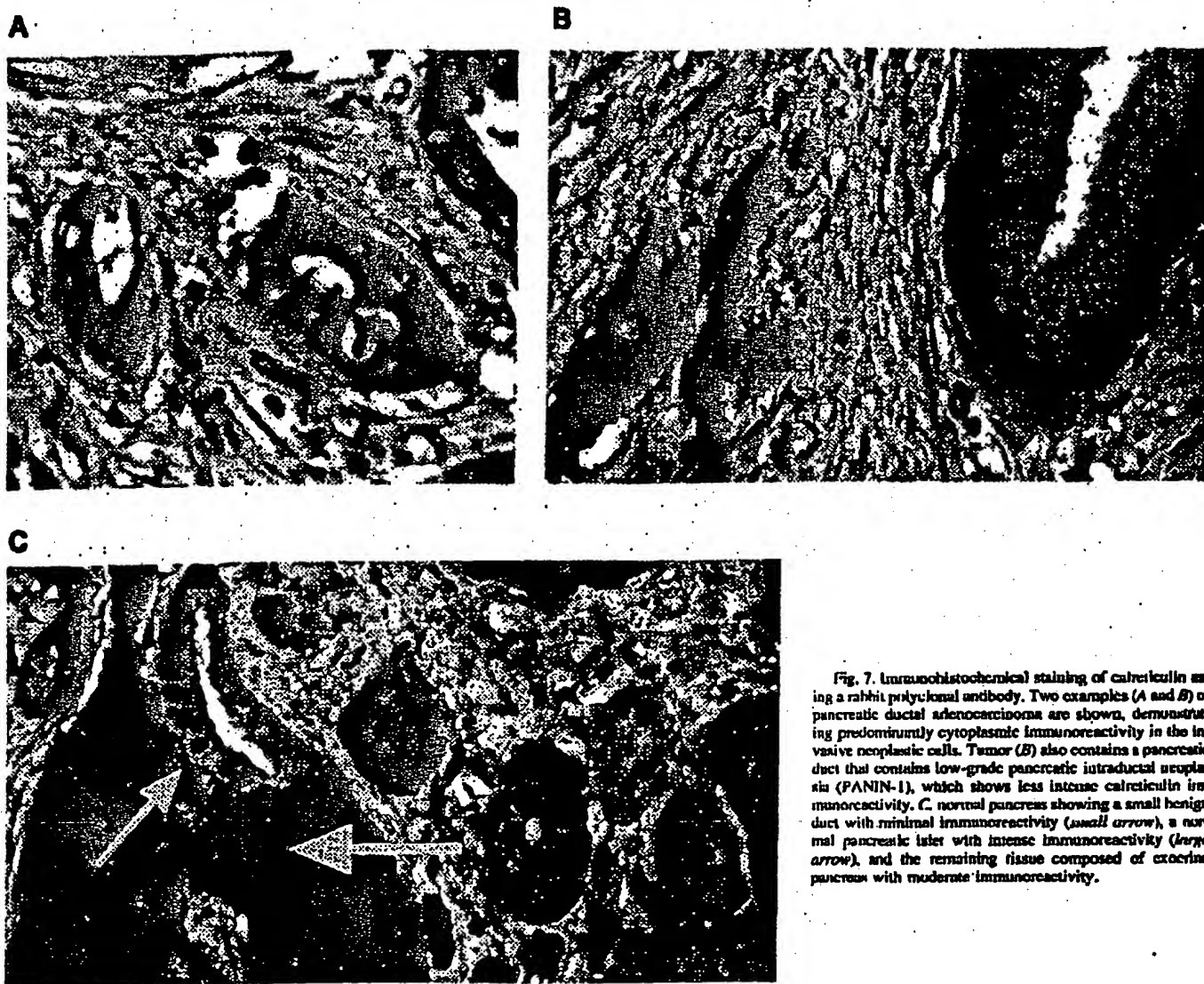


Fig. 7. Immunohistochemical staining of calreticulin using a rabbit polyclonal antibody. Two examples (A and B) of pancreatic ductal adenocarcinoma are shown, demonstrating predominantly cytoplasmic immunoreactivity in the invasive neoplastic cells. Tumor (B) also contains a pancreatic duct that contains low-grade pancreatic intraductal neoplasia (PANIN-1), which shows less intense calreticulin immunoreactivity. C, normal pancreas showing a small benign duct with minimal immunoreactivity (small arrow), a normal pancreatic islet with intense immunoreactivity (large arrow), and the remaining tissue composed of exocrine pancreas with moderate immunoreactivity.

level of protein expression was unlikely to contribute to the antigenicity of calreticulin.

Immunohistochemical Analysis of Calreticulin. Calreticulin expression in pancreatic and ampullary tumors was assessed by immunohistochemistry (Fig. 7, A and B), using a rabbit polyclonal anticalreticulin antibody and the pancreatic/ampullary tumor tissue array. Diffuse and consistent cytoplasmic immunoreactivity for calreticulin was observed in the majority of the pancreatic and ampullary adenocarcinomas. There were no significant staining differences with regard to tumor differentiation. Normal pancreatic ductal epithelium exhibited minimal reactivity, whereas normal pancreatic islets exhibited intense immunoreactivity, and normal exocrine pancreas exhibited moderate reactivity (Fig. 7C).

DISCUSSION

We have implemented a proteomics-based approach to identify proteins that elicit a humoral response in pancreatic cancer patients. This approach allows screening by Western blot analysis of patient sera for antibodies that react against separated tumor cell proteins. This study was focused on a search for autoantibodies to pancreatic tumor proteins present in the Panc-1 cancer cell line. We have shown

that a humoral response directed against calreticulin isoform 1 or 2, or both, occurred in 58.3% of pancreatic cancer patients. One of 18 chronic pancreatitis patients (5.6%) and 1 of 15 healthy controls (6.6%) demonstrated autoantibodies to calreticulin isoform 1; none demonstrated autoantibodies to isoform 2. None of the sera from patients with colon cancer exhibited reactivity against these proteins. One of 14 (7.1%) sera from lung adenocarcinoma patients demonstrated autoantibodies to calreticulin isoform 1. Two of 14 (14.3%) demonstrated autoantibodies to isoform 2.

Calreticulin is an abundant, high-capacity Ca^{2+} -binding protein found in the endoplasmic reticulum (ER) lumen of most cells of human origin. It has been shown to play a role in the regulation of a variety of cellular functions within the ER lumen (chaperone functions and Ca^{2+} storage and signaling) and calreticulin-dependent modulation of cell adhesion and gene expression at extra-ER sites (16). In particular, calreticulin interacts with *N*-linked oligosaccharides on nascent proteins in the ER lumen, with Ca^{2+} binding essential for this function.

It has been demonstrated that calreticulin elicits a humoral response in a variety of autoimmune diseases (17). Peptides transported into the lumen of the ER associate with calreticulin, as well as with protein

disulfide isomerase and gp96 (18). Moreover, calreticulin preparations purified from tumors elicit specific immunity to the tumor used as the source of calreticulin but not to an antigenically distinct tumor (19). This immunogenicity has been attributed to the peptides associated with the calreticulin molecule. The mechanism by which the calreticulin-peptide complex elicits immunity is unknown. A number of antigenic epitopes of calreticulin have been identified (20). The epitopes eliciting a humoral response in patients with autoimmune diseases have been reported to be located in the NH₂-terminal part of the molecule. Calreticulin is a component of the MHC class I peptide loading complex (21), and it has been demonstrated that calreticulin elicits tumor- and peptide-specific immunity (19). Interestingly, it has been shown that a particular form of calreticulin elicits a humoral response in hepatocellular carcinoma (22), with the reactive epitope occurring in a truncated form (CRT32, which includes the COOH-terminal portion), whereas the intact protein did not elicit reactivity. In our study, although the truncated form of calreticulin, CRT32, is present in the Panc-1 tumor cell line, it did not elicit immunoreactivity. This suggests that a specific mechanism of calreticulin processing may exist during carcinogenesis that may differ between tumor types.

A prerequisite for an immune response against a cellular protein is its presentation as an antigen. It is not clear why only a subset of patients with a specific tumor type develop a humoral response to a particular antigen. Immunogenicity may depend on the level of expression, posttranslational modification, or other types of processing of a protein, the extent of which may be variable among tumors of a similar type. We have demonstrated that calreticulin is not overexpressed in pancreatic tumor cell lines at either the mRNA or protein level, compared with lung, colon, or ovarian tumor cell lines in our study. Thus, the immunoreactivity of calreticulin is unlikely to be related to the level of protein expression. Furthermore, we were unable to demonstrate aberrant N-linked glycosylation of calreticulin in the pancreatic tumor cell lines (data not shown). It is possible that the antigenicity to the calreticulin isoforms may be arising from the aberrant expression of an unrelated protein in pancreatic cancer that generates an epitope that cross-reacts with calreticulin.

Although the calreticulin autoantibodies were largely restricted to patients with pancreatic cancer among the subject groups we investigated, additional studies are needed to determine the specificity of the calreticulin antibodies to pancreatic cancer. For example, although increased levels of calreticulin antibodies were found in pancreatic cancer, compared with chronic pancreatitis and other control groups, the relationship between tumor burden, tumor staging, and antibody levels needs additional clarification. Assessment of the utility of calreticulin autoantibodies as diagnostic markers in pancreatic cancer also needs to be addressed in additional studies. It is clear, however, that the proteomic approach that we have implemented, which allows for the screening of native proteins as they are expressed in tumor cells, has the potential to identify novel proteins that may have clinical utility in cancer.

ACKNOWLEDGMENTS

We thank Michele Le Blanc of the University of Michigan Tissue Core for excellent assistance with immunohistochemistry and Robert Hinderer for excellent technical assistance.

REFERENCES

- Brichory FM, Misk DE, Yim AM, et al. An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc Natl Acad Sci USA* 2001;98:9824-9.
- Le Naour F, Misk D, Krause M, et al. Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clin Cancer Res* 2001;7:3328-35.
- Gure AO, Akorki NK, Stockert E, Scanlan MJ, Old LJ, Chen YT. Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. *Cancer Res* 1998;58:1034-41.
- Stockert E, Jager E, Chen YT, et al. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J Exp Med* 1998;187:1349-54.
- Jemal A, Murray T, Samuels A, Ghafoor A, Ward E, Thun MJ. Cancer statistics, 2003. *CA-Cancer J Clin* 2003;53:5-26.
- Hatanaka Y, Suehiro Y, Fukui M, Shikichi K, Inai K, Hinoda Y. Circulating anti-MUC1 IgG antibodies as a favorable prognostic factor for pancreatic cancer. *Int J Cancer* 2003;103:97-100.
- Raele J, Oremek G, Welker M, Roth WK, Caspary WF, Zenzon S. p53 autoantibodies in patients with pancreatitis and pancreatic carcinoma. *Pancreas* 1996;13:241-6.
- Maschke H, Hundermark C, Miska S, Voss M, Kalhoff H, Sureschwer HW. Autoantibodies in sera of pancreatic cancer patients identify recombinant factor Rad51 as a tumour-associated antigen. *J Cancer Res Clin Oncol* 2002;128:219-22.
- Kouonon J, Bubendorf L, Kalliolehti A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844-7.
- Maschke H, Jost K, Opitz S, et al. DNA repair and recombination factor Rad51 is over-expressed in human pancreatic adenocarcinoma. *Oncogene* 2000;19:2791-5.
- Abu-Shakra M, Boskila D, Ehrenfeld M, Conrad K, Shoenfeld Y. Cancer and autoimmunity: autoimmune and rheumatic features in patients with malignancies. *Ann Rheum Dis* 2001;60:433-41.
- Kubic R, Skolnick MM, Neel JV, Hanash SM. A two-dimensional electrophoresis-related laboratory information processing system: spot matching. *Electrophoresis* 1991;12:736-46.
- Gharahdaghi F, Weinberg CR, Mergler DA, Imai BS, Hanash SM. Mass spectrometric identification of proteins from silver-stained polyacrylamide gels: a method for the removal of silver ions to enhance sensitivity. *Electrophoresis* 1999;20:601-5.
- Giordano TJ, Shadden KA, Schwartz DR, et al. Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. *Am J Pathol* 2001;159:1231-8.
- Rickman DS, Bobek MP, Misk DE, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* 2001;61:6883-91.
- Krause KH, Michalak M. Calreticulin. *Cell* 1997;88:439-43.
- Eggleston P, Ward FJ, Johnson S, et al. Fine specificity of autoantibodies to calreticulin: epitope mapping and characterization. *Clin Exp Immunol* 2000;120:384-91.
- Spee P, Neefjes J. TAP-translocated peptides specifically bind proteins in the endoplasmic reticulum, including gp96, protein disulfide isomerase and calreticulin. *Eur J Immunol* 1997;27:2441-9.
- Basu S, Srivastava PK. Calreticulin, a peptide-binding chaperone of the endoplasmic reticulum, elicits tumor- and peptide-specific immunity. *J Exp Med* 1999;189:797-802.
- Lieu T-S, Newkirk MM, Capra JD, Sotthearner RD. Molecular characterization of human Ro/SS-A antigen: amino terminal sequence of the protein moiety of human Ro/SS-A antigen and immunological activity of a corresponding synthetic peptide. *J Clin Invest* 1988;82:96-101.
- Michalak M, Corbett EH, Mesaeri N, Nakamura K, Opas M. Calreticulin: one protein, one gene, many functions. *Biochem J* 1999;344:281-92.
- Le Naour F, Brichory F, Misk DE, Brechet C, Hanash SM, Beretta L. A distinct repertoire of autoantibodies in hepatocellular carcinoma identified by proteomic analysis. *Mol Cell Proteomics* 2002;1:197-203.

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☒ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.